The Dissertation Committee for Martin Blom
certifies that this is the approved version of the following dissertation:

Automated Prediction of Human Disease Genes

Committee:

Edward M. Marcotte, Supervisor

Inderjit S. Dhillon

Oscar Gonzalez

William H. Press

Claus O. Wilke

Automated Prediction of Human Disease Genes

by

Martin Blom, Civilingenjörsexamen

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2012

Dedicated to my wife Amrita.

Acknowledgments

There are so many people that have helped make this dissertation happen. First and foremost, of course, I want to thank Edward Marcotte for being a fantastic boss, for his constant enthusiasm and curiosity, and for creating an amazing place to work and learn.

I want to thank all my friends and colleagues in the Marcotte lab. You guys have made the lab a wonderful place to be over the past few years. Special thanks go to Kris McGary, Traver Hart, Peggy Wang, Blake Borgeson, and Taejoon Kwon for being excellent sounding boards over the years, to Jeremy O'Connell, Jagannath Swaminathan, Zhihua Lee, Christine Vogel and Dan Boutz for keeping the rest of us grounded, and to Mark Tsechansky for making the lab such a peaceful and conflict free work environment.

Thanks are of course due to my many collaborators. Insuk Lee, John Woods, Nagarajan Natarajan, and Ambuj Tewari – it's been a pleasure.

I want to thank the members of my committee: Professors Oscar Gonzalez, Claus Wilke, Inderjit Dhillon, and Bill Press. Special thanks go to Inderjit Dhillon and Bill Press who taught wonderful courses in statistics and machine learning, and taught me most I know about how to deal with data.

I would also like to thank all my friends in Austin who have made my time outside of lab such a blast. Thanks Larsson and Jesper, for the *nubbe* and songs, and

for luring me over to biology in the first place. Thank you Robert for the squash. Thank you Alex, for, well, actually, many many pleasant discussions. Thank you Benni for the gin and the tonic and uncountable dinners.

Special thanks are due to my dear family, who supported my strange decision to leave the nurturing care of the socialist mothership and move to Texas.

Finally, thank you Amrita. You would have been reason enough to go to Texas.

Automated Prediction of Human Disease Genes

Publication No. _____

Martin Blom, Ph.D.
The University of Texas at Austin, 2012

Supervisor: Edward M. Marcotte

The completion of the human genome project has led to a flood of new genetic data, that has proved surprisingly hard to interpret. Network "guilt by association" (GBA) is a proven approach for identifying novel disease genes based on the observation that similar mutational phenotypes arise from functionally related genes.

However, GBA has been shown to work poorly in genome-wide association studies (GWAS), where many genes are somewhat implicated, but few are known with very high certainty. In the first part of this work, I resolve this by explicitly modeling the uncertainty of the associations and incorporating the uncertainty for the seed set into the GBA framework. I demonstrate a significant boost in the power to detect validated candidate genes for Crohns disease and type 2 diabetes by comparing the predictions from my method to results from follow-up meta-analyses, with incorporation of the network serving to highlight the *JAK–STAT* pathway and associated adaptors *GRB2/SHC1* in Crohns disease and *BACH2* in type 2 diabetes. Consideration of the network during GWAS thus conveys some of

the benefits of enrolling more participants in the GWAS study. More generally, we demonstrate that a functional network of human genes provides a valuable statistical framework for prioritizing candidate disease genes in GWAS-based studies.

Furthermore, functional gene networks are not the only kind of information that can be used to predict gene–phenotype associations. In the second part of this thesis, I show that gene–phenotype associations in model species from species as distantly related to humans as *E. coli* is another valuable source of information, that can be mined using methods similar to those used in recommender systems.

Finally, in the last part of this thesis, I present a machine learning formalism that combines the functional gene network and model species phenotype information. I show that this approach outperforms the state of the art methods for gene–phenotype association prediction using cross-validation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Background – The Genetics of Human Disease

1.1 Introduction

The completion of the Human Genome Project [1, 2] and the HapMap project [3] gave us a map of common human genetic variation that has enabled an incredible growth in our understanding of the genetics of many human diseases over the last decade. Genome-wide association studies (GWAS), which assay hundreds of thousands of mutations in thousands of patients, have identified risk alleles for many common diseases (see, for example [4]), and exome sequencing enables unbiased discovery of rare disease-causing mutations, such as the discovery of the causal gene for the rare Mendelian disorder Miller syndrome [5].

Meanwhile, our growing understanding of genetics and molecular biology has enabled an alternative approach to the genome-wide, unbiased techniques epitomized by GWAS and exome sequencing. Sequencing of candidate genes have identified mutations underlying many diseases, such as Rett syndrome [6], and ever better model systems for studying human diseases have been developed in mouse, fly, fish, and many other model species. A very successful example of this approach are so-called "guilt-by-association" methods, where new candidate genes are predicted for diseases by their association to already known causal genes for the disease. This association can be of many forms, such as protein–protein

interactions, co-expression, genetic interactions, and so on.

One particular method that has been pursued by this lab, is to use an integrative Bayesian approach to encode many different kinds of functional associations into a *genome-wide functional gene–gene interaction network*, first developed for baker's yeast by Edward Marcotte [7] and Insuk Lee [8]. These functional gene networks have been shown to be effective tools for finding new genes involved in, for example, gene loss-of-function phenotypes in yeast [9], and RNAi knockout phenotypes in nematodes [10]. In this work, we show that GBA using HumanNet, a new functional gene–gene interaction network for humans, can also be used to find new human disease genes. In particular, I will present a method for using network GBA when the labels are "fuzzy", that is, when we have a large number of genes that are *likely* to be involved in a disease, but not certain. This allows us to bring network-guided GBA to the GWAS setting, and identify pathways and mechanisms likely to be involved in Crohn's disease and type 2 diabetes.

Another form of association that has been used successfully to identify disease genes in human disease is the *phenolog* approach proposed by Kris McGary *et al.* [11], who identify pairs of phenotypes (which they call phenologs) in different organism that have an unusually high number of genes in common. They hypothesize that this etiological overlap is due to some kind of evolutionarily conserved mechanism, and that the genes involved in only one of the phenotypes in the phenolog pair should therefore be good candidate genes for the other one. In this work, I extend this framework to draw on multiple distantly related phenotypes for predicting new genes for human diseases.

Finally, to conclude this dissertation, I will present a machine learning framework that combines HumanNet and the phenolog idea into a single, unified prediction method. I will show the power of this combined framework for predicting gene – disease associations, and use it to highlight a weakness in one of the schemes commonly used to test the performance of disease gene prediction methods.

To set the context for this work, I will first give a very brief overview of the molecular biology and genetics concepts that are necessary to understand this work. I will then present the results in three chapters, one chapter on HumanNet and GWAS, one chapter on predicting disease genes using multiple phenotypes from distantly related species, and one chapter on how to combine HumanNet and multiple species, and on how these results should be tested.

1.2 Some molecular biology

Since Watson and Crick [12], what is known as the *central dogma of molecular biology* has held that there are three classes of information bearing biopolymers, DNA, RNA, and proteins, and that information can flow between these in nine possible ways [13], of which three are general (normal), three special (unusual) and three unknown (never observed).

DNA (deoxyribonucleic acid) and RNA (ribonucleic acid) are both large molecules, made up of long chains of nucleotides. For DNA, these are adenine (A), cytosine (C), thymine (T) and guanine (G). For RNA, uracil (U) is used instead of thymine. These nucleotides come in two parings: A-(T/U) and C-G. In DNA, two chains (called strands) combine into a double helix with a perfect complement, so

that across from each A in one chain there is a complementary T in the other, and across from each C there's a G.

During transcription, molecules called RNA polymerase transcribe some short set of instructions written by the DNA into a RNA molecule called a messenger RNA (or mRNA for short). The mRNA is then translated into a string of amino acids called a *protein* by molecular machines called ribosomes. The proteins are involved in most of the actual work in the cell — e.g. as enzymes that catalyze different reactions, as scaffolds in different structures in the cell, and by binding to RNA or DNA. A stretch of DNA that codes for a specific protein is often called a *gene*.[1]

1.3 Interactions and pathways

Proteins often interact with each other in different ways, either transiently (e.g. by phosphorylating a different protein) or on a more long term basis (e.g. by forming a long-lived complex with another protein). They are also involved in the regulation of protein production in many ways, e.g. by binding to DNA to up or down regulate the expression of genes. Often, interactions of different kinds are strung together into complex regulatory mechanisms called *pathways*, which can be involve either protein-protein interactions (such as phosphorylation cascades and other signaling mechanisms) or many levels of protein-DNA interactions (such as

[1]This is a rather imprecise use of the term. Geneticists define a gene to be a unit of inheritance, which could involve any kind of structure in the DNA. Thus, for many traits the stretch of DNA that regulates the trait does not code for any other molecule at all, but might instead be involved in the regulation of the expression of a protein coding gene. I will, however, follow the standard convention of the field and use the term sloppily.

when a transcription factor up regulates the expression of another transcription factor), or combinations of the two. In this work, I will refer to both of these kinds of interactions as *functional gene-gene interactions*.

1.4 Haplotypes, and the organization of DNA

When DNA gets copied, sometimes mistakes slip in which are then inherited by all successive copies. Such mistakes are called mutations, and can both increase and decrease the fitness of the organism — sometimes depending on the environment the organism finds itself in.

DNA is organized into large molecules called *chromosomes*, and every human cell (except for germ line cells) carries two similar, but not identical copies of each chromosome — one inherited from the father, and one from the mother. Cells that carry two copies of each chromosome are called *diploid*, and cells that carry a single copy of each chromosome are called *haploid*.

During sexual reproduction, cells undergo a special kind of cell division called meiosis, in which a diploid mother cell produces haploid daughters. In men, these daughter cells are sperm cells, in women, egg cells. During meiosis, the DNA in each chromosome recombines with the corresponding stretch of DNA in the other copy of the same chromosome, which causes the chromosomes in the daughter cells to be a mix of the chromosomes in the mother cell, with alternating stretches of DNA derived either from the paternally or the maternally inherited chromosome.

Because of this crossover process, if two variants exist in an individual, they

are likely to be inherited together if they are located close to each other on the actual chromosome. If a carrier of such a close pair is reproductively successful, a large fraction of the population might end up carrying both mutations together. These blocks of co-inherited genetic variants are called *haplotypes*. As generations go by, the linkage between the variants in a haplotype block will get weaker and weaker due to repeated crossover. Therefore, the linkage between different loci will be weaker the older a population is. We therefore have a clearer haplotype structure in e.g. European and East Asian populations (that have undergone quite recent population expansions during the agricultural revolution) than in most aboriginal African populations, where the haplotypes have been mixed up through repeated mating.

1.5 Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs, pronounced "snips") are the most common kind of genetic polymorphism in terms of the number of occurrences [14]. In a SNP, a mutation has occurred at a certain position in the genome that changes one letter of the genetic alphabet into another one. This variant has then spread out into the general population, and now exists at a certain percentage in the population studied, while the original makes up for the rest. We might not right now know which of these two variants is the original and which is newer, but we can measure which one is more common, and which one is rarer. The more common version is called the major allele, and the less common the minor allele.

After the completion of the Human Genome Project [1, 2], one of the major

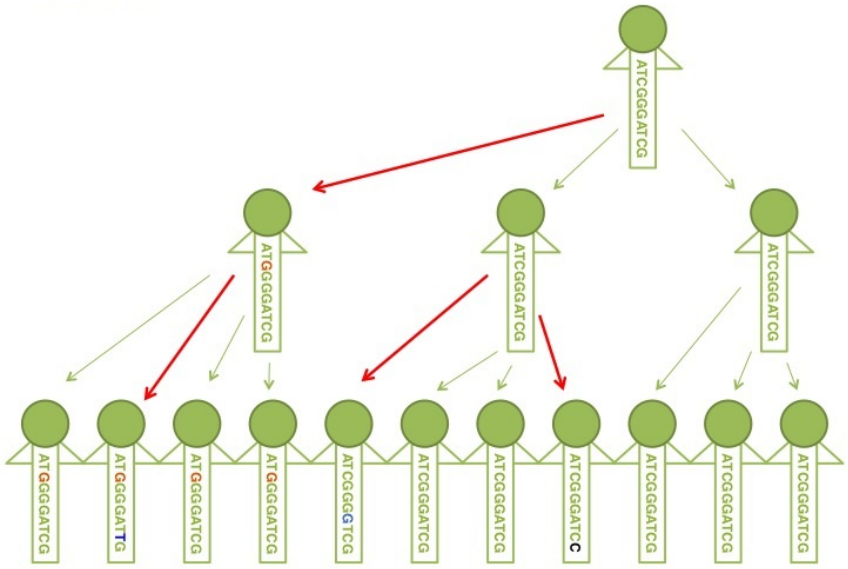


Figure 1.1: **Population structure arises from mutations in ancestors to the present population.** Here we show how mutations propagate in a haploid population. Parent – offspring relationships are shown with arrows, and mutated nucleotides are shown in a contrasting color. Haplotypes arise when an individual sees considerable reproductive success, like the leftmost individual in the second generation above; all its offspring share the same $C \rightarrow G$ mutation, which defines the haplotype.

endeavors in human genetics has been the mapping of the common genetic variations in human beings, the so-called International HapMap Project [3]. HapMap has identified millions of common SNPs (minor allele frequency (MAF) $> 5\%$), and also found which ones were strongly correlated. This allowed the construction of a type of device called a *SNP chip* (reviewed in [15]).

A SNP chip is a microchip-like little device with about a million dots on it. Each dot contains a piece of DNA also found on the genome, centered around one of the common SNPs. For each SNP there are at least two dots, one for each allele. By measuring which version a person's DNA binds to, it is possible to determine if the person is homozygous for the major version, the minor version or heterozygous at that SNP. Since even early versions of these chips contained more than a million such dots, one can easily measure hundreds of thousands of SNPs in a single measurement. This, and the fact that each chip only costs a few hundred dollars to make, has made large-scale measurement of genetic variations possible.

1.6 Genome-wide association studies, and the common disease — common variant hypothesis

Many common disease — for example diabetes, Crohn's disease, obesity and bipolar disorder — have strong hereditary components. In the years following the completion of the human genome project, there was a lot of uncertainty over how the genetic architecture of this hereditary component might be structured. Two common schools of thought emerged: One, the common disease – common variant (CDCV) hypothesis, and the common disease – rare variant (CDRV) hy-

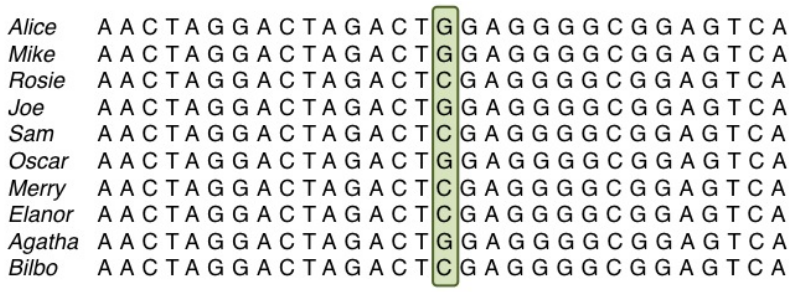
pothesis [16, 17]. The CDCV hypothesis posited that all humans harbored a large number of genetic variants that each had a very minor effect on the disease risk. Since no single variant significantly increased the risk that the bearer would develop the disease, there was no strong selection against them. Only when someone, by bad luck in the great lottery known as sexual mating, happens to draw many more of the risk variants than most people, do they develop the disease.

The CDRV hypothesis on the other hand holds that mutations that increase the risk for disease would be selected against, and we wouldn't see many of them. Instead, it hypothesized that the common diseases we see are caused by many different mutations, that spontaneously arise every now and then. If they arise in a type 1 diabetes gene, the unlucky bearer and his offspring are more likely to develop diabetes, and if they occur in a bipolar disorder gene, the bearer is likely to develop bipolar disorder. Since there is a selective pressure against having these diseases, they will usually only persist for a few generations before they go extinct. In this scenario, any single variant will be much rarer than the ones hypothesized by the CDCV hypothesis, but will have a much stronger effect on the unfortunate individual who happens to carry them.

With the advent of the SNP chips described above, we were given the tools to experimentally test the CDCV hypothesis. A type of study, called a Genome Wide Association Study (GWAS) was devised [15]. Most GWAS are planned using a case-control design. In a case-control study, about half of the participants have developed the disease (the cases) and the other half are healthy (the controls). Each participant's genetic make-up is measured using a SNP chip that is designed to

be as unbiased as possible in the SNPs it measures, and the cases and controls are compared. If the distribution of a SNP differs between the cases and the controls, it is likely that some genetic variant in the vicinity of this SNP is causing people to develop the disease. For a more complete overview of GWAS, see the reviews in [18] or [19].

The outcome of these GWAS has met mixed reception. On the one hand, they have identified hundreds of new gene associations for a wide range of disorders. On the other hand, they can only explain a very small fraction of the known heritability of these diseases. This has led to widespread debate over where this "missing heritability" might be hidden [20]. Could it be in non-additive interactions between the different loci, so called *epistatic* interactions? Is it because the diseases themselves are poorly characterized? Or is the CDRV hypothesis partially true, and rare, but strong, mutations account for the heritability that is "missing"? For now, the jury is still out, but there is probably some partial truth to all of the above explanations.



	GG	GC	CC
Short	43	633	2309
Tall	130	993	1881

$$\chi^2 = \sum \frac{(obs - exp)^2}{exp}$$

Figure 1.2: **A C/G SNP in a haploid population.** Note that real human populations have two copies of each gene. In a GWAS, this SNP is analyzed to see if it's distribution differs significantly between the case and control population. Since a GWAS analyzes millions of SNPs, the p-value cutoff needed to be considered significant is very strict.

Chapter 2

Prioritizing candidate disease genes by network-based boosting of genome-wide association data

The first major part of my work makes use of HumanNet to improve the predictions of GWAS. This chapter is adapted from the paper "Prioritizing candidate disease genes by network-based boosting of genome-wide association data", which I published with Insuk Lee as joint first author in Genome Research, 2011 [21].

In this paper, Insuk Lee was responsible for the construction of the network and the basic analyses of how phenotypes can be predicted using the network, Peggy Wang compared different network prediction algorithms, and I adapted the network GBA strategies to the GWAS setting. Insuk, Edward and I wrote the paper together.

2.1 Introduction

Causal genes remain extraordinarily difficult to identify in most genetic diseases, and in particular, in highly polygenic disorders, for which current approaches are most limited [22], identifying causal genes is a major barrier to progress in understanding these diseases. More generally, traditional linkage analyses have mapped causal genes for many diseases, often using positional cloning, but these

methods are difficult and time-consuming [23]. However, genome-wide association studies (GWAS) have opened the way to unbiased discovery of large numbers of disease genes in a more efficient manner.

A typical GWAS analysis involves comparing case and control individuals at selected single nucleotide polymorphisms (SNPs) or, more recently, copy number variants (CNVs). SNPs representing common haplotype blocks are measured genome-wide (at approx. 500,000 1,000,000 locations), and the disease-associated genetic markers are identified (reviewed in [24]). The SNPs that show association strong enough to surpass a genome- wide significance threshold are then analyzed for chromosomal proximity to genes that might cause the disease, or otherwise affect its etiology. However, even though the data from GWAS support a great number of loci involved in common diseases, it is hard to separate many of the causal genes from the background noise of the hundreds of thousands of SNPs in the assay. Consequently, GWAS suffer from a lack of statistical strength, requiring large test populations to overcome the large multiple hypothesis correction needed in evaluating hundreds of thousands of candidate loci.

The lack of sufficient statistical power forces GWAS studies to ignore weaker loci, focus primarily only on the strongest genetic effectors, and genotype thousands of individuals (e.g., [4]). Moreover, the combinatorial effects of multiple disease genes are often not simply additive but epistatic [15, 25, 26], further hampering their discovery. Simply considering pairs of interacting loci increases the strength of associations required by orders of magnitude so as to be able to overcome the multiple testing criteria, requiring tens of thousands of individuals [27]. Rarely has

genetic association to allele triplets (or higher) been examined by these or any other approaches. Linear additive models have been successfully built, most notably for 54 alleles useful for predicting human height [28–30], one of the first quantitative human traits successfully addressed to this degree. Finding these alleles nonetheless required genotyping 63,000 individuals over the course of 3 studies, each explaining ¡4% of the variance in height. Recent analysis of approx. 300,000 SNPs, without regard to the significance of their association, demonstrated that a total of 45% of the variance in height could potentially be explained, with most effects too small to pass significance tests [31].

However, the polygenic nature of a disease may also offer potential opportunities to more efficiently discover new and relevant genes. In particular, we might expect that the genes associated with a disease will often organize into pathways or functional groupings linked to the disease formation and progression. Thus, knowing some disease genes in advance, it may occasionally be possible to apply guilt-by-association (GBA) in gene networks (reviewed in [32]). In particular, it is now possible to construct large gene network models, as has been done e.g. for yeast, worms, plants, mice, and humans, summarizing thousands of functional associations among genes, as reviewed in [33–36]. Gene pairs are coupled in these networks if they are inferred to participate in the same biological process [7] and may have corresponding measures of confidence [8, 37–40]. GBA in such networks has been shown to correctly identify disease and phenotype-linked genes based on their network connections to previously known genes (e.g., [9, 10, 41–46]), based on the observation that genes involved in a common biological process often tend be

associated with similar mutational phenotypes, as seen e.g. in [10, 41, 42, 47].

In principle, the GWAS-based association of genetic loci with a disease and the functional association of genes into pathways represent independent sets of observations that can be logically combined to improve identification of relevant disease genes. For example, networks have been applied to search for interacting loci in human GWAS data [48, 49] and in yeast [50], to identify GWAS- and cancer genome-enriched pathways [51, 52], and to rank genes in implicated chromosomal intervals [53–55]. Other studies have looked at previously studied pathways for a disease, and tried to improve the ranking of the candidate genes using this information (e.g. [56, 57]; more reviewed in [58]). Here, we have tested and expanded the general validity of the approach of using functional networks for prioritizing candidate disease genes. We propose a theoretical framework for combining the large-scale, unbiased pathway and association information encoded by functional gene networks and GWAS studies respectively, showing improvements in performance as judged by data from GWAS meta-analyses.

First, we describe the construction of a functional network for human genes. This network spans 87% of validated protein coding genes, and provides strong predictive power for a majority of currently known genetic diseases. We evaluate six alternate approaches for prioritizing candidate disease genes using this network, and demonstrate the strongest overall performance with algorithms related to Googles PageRank. We then show that this network, in conjunction with genome-wide association data for Type 2 diabetes and Crohns disease, boosts the identification of disease-associated genes that were discovered in later meta-analyses.

This work suggests both a specific strategy and a general path to future improvements for the interpretation of GWAS data. Taken together, our work demonstrates that a high-quality functional network for human genes can provide a powerful resource for identifying causal genes in human disease.

2.2 Results

HumanNet: an extended functional gene network for H. sapiens

To test the ability of functional networks to improve gene association studies, we first constructed a genome-scale functional network for human genes. Diverse distinct lines of evidence, spanning human mRNA co-expression, protein-protein interactions, protein complex, and comparative genomics datasets, in combination with similar lines of evidence from orthologs in yeast, fly, and worm, were analyzed, using an approach previously developed and validated for yeast [8, 59], C. elegans [10, 60] and Arabidopsis [46]. In total, 21 large-scale genomics and proteomics datasets from the 4 species (see Methods, Tables S1 and S2) were integrated into a functional gene network spanning 476,399 scored functional couplings between 16,243 (87%) of validated human protein encoding genes (Figure 2.1).

HumanNet predicts cellular loss-of-function phenotypes

To evaluate the predictive power of the new network, we first examined cellular-level phenotypes. Many human diseases reflect failures of core cellular machinery, e.g. failures of metabolism, DNA repair, replication, etc. For example, hereditary nonpolyposis colorectal carcinoma arises from mutations in DNA

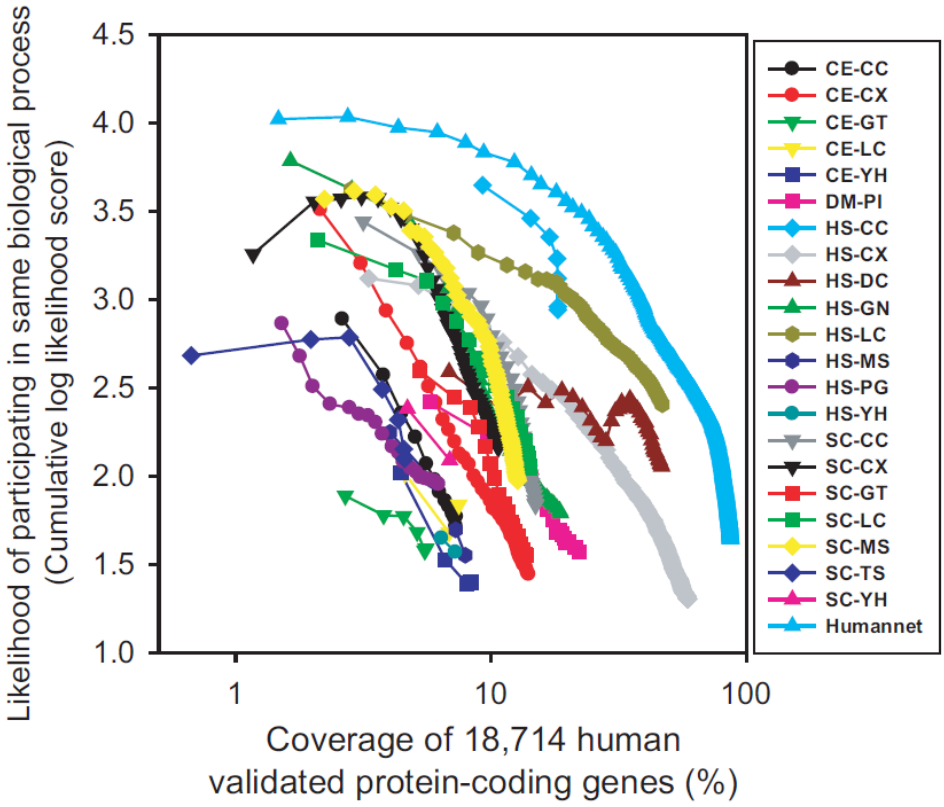


Figure 2.1: **Construction of a genome-scale human gene network, HumanNet.** 21 diverse functional genomic and proteomic data sets (Table 1) were evaluated for their tendencies to link genes in the same biological processes. Pairwise gene linkages derived from the individual datasets were then integrated into a composite network of higher accuracy and genome coverage than any individual data set. The integrated network (HumanNet) contains 476,399 functional linkages among 16,243 (86.7%) of the 18,714 genes encoding validated human proteins. The plot x axis indicates the log-scale percentage of the 18,714 genes covered by functional linkages derived from the indicated datasets (curves); the y axis indicates the predictive quality of the datasets, measured as the cumulative log likelihood of linked genes to share Gene Ontology (GO) biological process annotations, tested using 0.632 bootstrapping and plotted for successive bins of 1,000 linkages each (symbols). Data sets are named as XX-YY, where XX indicates species of data origin (CE, *C. elegans*; DM, *D. melanogaster*; HS, *H. sapiens*; SC, *S. cerevisiae*) and YY indicates data type (CC, cocitation; CX, mRNA coexpression; DC, domain cooccurrence; GN, gene neighbor; GT, genetic interaction; LC, literature-curated protein interactions; MS, affinity purification/mass spectrometry; PG, phylogenetic profiles; PI, fly protein interactions; TS, tertiary structure; and YH, yeast two-hybrid). Detailed descriptions are listed in Table S1.

Table 2.1: **Selected top-ranked Crohns disease and type 2 diabetes genes** for which network data added support to GWAS evidence, measured as an increase in odds (prior=1.7 for each).

Crohn's disease

Gene name	New rank	Original rank	Log odds increase	Interaction partners
NOD2	1	1	0	
ATG16L1	2	2	0.53	*CAPN9*
IL23R	3	3	0.76	*STAT3*
CYLD	4	4	0.52	*TRAIP*
PTPN2	5	6	0.76	*STAT3*
GRB2	7	99	3.63	*DAG1, APP, STAT3, DDK1, PPP2R2B*
STAT3	8	17	1.88	*IL23R, PTPN2, GRB2*
BSN	9	9	0.61	*CAMKV, ERC2*
DAG1	11	21	1.6	*TCTA, GRB2*
PPM1K	16	125	2.27	*CDK14, CAMKV, CLK3, MAGI2*
SHC1	17	6125	3.98	*PTPN2, STAT3, DOK1, GRB2, USP4,PTPN2, PPM1K*
SRC	20	11633	4.38	*MAGI2, DAG1, STAT3, GRB2, USP4, PTPN2, PPM1K*
CAPN9	22	18	0.58	*ATG16L1*
TRAIP	28	327	1.91	*BATF, CREM, CYLD, TRAIP,USP7*
JAK2	38	3139	2.95	*IL23R, STAT3, GRB2, IL12RB2, PPM1K, MAGI2*

Type 2 diabetes

Gene name	New rank	Original rank	Log odds increase	Interaction partners
TCF7L2	1	1	0	
THBS2	2	5	0.36	*ISLR*
CDKAL1	3	2	0	
TSPAN8	4	3	0	
PARD3B	10	13	0.22	*KIF23*
KIF23	14	44	1.05	*MELK, FAM49A, DYNCH1H1, GTSE1, PARD3B*
FAM49A	16	42	0.9	*ANKS1B, KIF23, ANKS1A*
ISLR	17	26	0.49	*THBS2, ZNF532*
BACH2	18	200	1.66	*TCF7L2, PARD3B, CREB5*
ANKS1A	23	30	0.32	*FAM49A*
XYLB	27	34	0.36	*ATG7*
MAGI2	29	65	0.67	*ALK, CHUK, PRKG1, MELK, DYRK1A*
CDC42	35	191	1.18	*PARD3B, ATG7*
MELK	38	51	0.46	*MAGI2, KIF23*
CTNNB1	76	3099	1.88	*ATG7, TCF7L2, LOH12CR1, CHUK, MAGI2*

mismatch repair [61, 62], Zellweger syndrome arises from mutations in peroxisome biogenesis [63], and leukoencephalopathy with vanishing white matter arises from mutations in any of the subunits of translation initiation factor eIF2B [64, 65]. A network for even a single eukaryotic cell will capture many of these basic processes, and has the potential to prove predictive for genes for diverse human diseases. We therefore investigated if the human gene network was predictive of cellular-level mutational phenotypes, focusing on cell survival and proliferation phenotypes from loss-of-function studies in cell culture.

We first asked if genes essential to cell viability could be accurately identified using the gene network. Schlabach and colleagues identified about 600 genes that affect the viability and proliferation of normal human mammary epithelial cells (HMEC) by using multiplex short hairpin RNA (shRNA) screening [66]. Although assayed largely for proliferation defects, these genes are highly likely to be essential for HMEC cell growth, given the incompletely penetrant phenotype induced by shRNA knockdown [67]. We found that the essential HMEC genes were, indeed, highly connected in HumanNet (Figure 2.2A), as assessed by cross-validated receiver operating characteristic (ROC) analysis (see Methods). For example, about 18% of all known essential genes, but only 2% of all genes not known to be essential, are connected to known essential genes in HumanNet, a nine-fold enrichment. From these results we conclude that essential genes can be predicted on the basis of their connectivity to other essential genes in HumanNet.

This general level of predictability was also observed for more specific cellular phenotypes. We tested if genes known to be required for HIV infection,

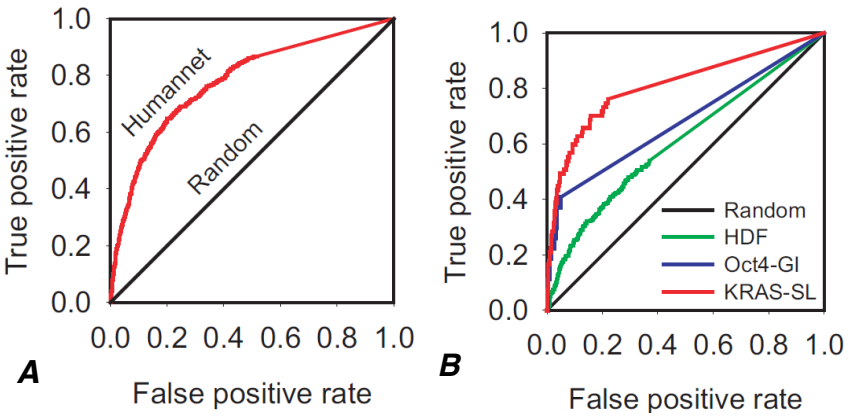


Figure 2.2: **Genes associated with many phenotypes are highly connected in HumanNet.** **A** Essential genes were highly interconnected in HumanNet, and thus predictable from the network, as shown by ROC analysis. Genes were ranked by their sum of network edge weights to the known essential genes, measuring recovery of known essential genes (true positives) and other genes (false positives) using leave-one-out cross-validation. **B** Specific phenotypes are predictable by HumanNet. Genes involved in more specific cellular phenotypes – host factors required for HIV infection (HDF) [68], modulators of *OCT4* (also known as *POU5F1*) expression (Oct4-GI) [69], and synthetic lethal partners of activated *KRAS* alleles (*KRAS*-SL) (Luo *et al.* 2009) – were also well predicted by their interconnectivity in HumanNet, calculated as for **A**.

as measured by large-scale RNAi knockdown [68], were predictable by guilt-by-association in HumanNet. Indeed, they showed a moderate degree of predictivity, at a level significantly higher than random chance (Figure 2.2B).

The essentiality and viral infectivity phenotypes described above are single gene phenotypes, but yeast and worm gene networks have also proven generally predictive for bigenic phenotypes, such as synthetic genetic interactions (e.g., [60]). We therefore next asked if the human gene network could predict genetic interactions, focusing on two large-scale RNAi screens performed in mammalian cell culture. The first screen identified genes modulating expression of a core stemness regulator *Oct4* in mouse embryonic stem cells [69]. The second found genes acting as synthetic lethal interaction partners with oncogenic *KRAS* mutants expressed in a colorectal cancer cell line, screening for genes whose knockdown in the activated *KRAS* background resulted in cellular lethality [70]. In both cases, genes identified by the screens were well-predicted by guilt-by-association in HumanNet at rates significantly higher than random expectation (Figure 2.2B). The high predictive strength (AUC = 0.81) for *KRAS* interactors is particularly notable, as such genes might be useful as cancer cell specific drug targets [70]. More generally, these tests confirm that the human gene network is predictive of a variety of cellular level loss-of-function phenotypes, including specific bigenic traits.

Genes linked to specific mouse mutational phenotypes and human diseases are predictable by guilt-by-association in the network

The cellular-level results demonstrate that genes for cell viability and proliferation phenotypes can be identified based on network connectivity in HumanNet.

A further trend for genes linked in the network to share tissue-specific expression patterns (Figure 2.3) implies that the network could potentially predict more specific organism-level mutational phenotypes as well. This notion has previously been explored for human diseases by considering network connections among known disease genes, prioritizing the genes most highly connected to the known causal genes as being likely new candidate genes for that disease [9, 41, 42, 45], as illustrated in Figure 2.4. Such approaches primarily consider direct network connections to known disease genes, but related work on predicting gene function from networks (reviewed in [71, 72]) has shown wide benefits of also appropriately considering indirect network connections (e.g., as in [73]), and tests have confirmed the utility of these so-called network diffusion algorithms for predicting RNAi phenotypes in worms and loss-of-function phenotypes in yeast cells [74]. Here, we implemented a representative set of both types of algorithms, collectively termed label propagation algorithms and chosen by their successful application in yeast and worm networks [74], for inferring disease genes based on network connectivity, evaluating them for their overall predictive ability using cross-validation and ROC analysis.

Specifically, we considered six methods of network label propagation. The first are two methods that consider only direct network neighbors: (1) neighbor counting [76], in which the genes with the most neighbors already linked to the disease are most highly scored, and (2) naïve Bayes label propagation, in which the sum of the HumanNet linkages to implicated neighbors is used rather than their count [59], corresponding to the naïve Bayes estimate for a gene to partici-

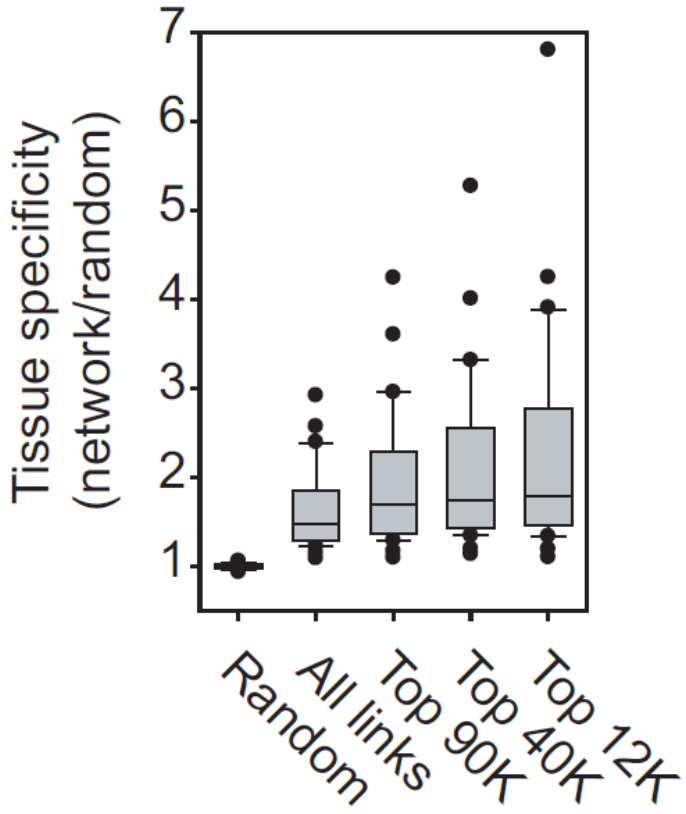


Figure 2.3: **Network-linked gene pairs were substantially more likely to show similar tissue specificity in their expression patterns.** This was measured as the likelihood of co-occurrence of transcripts of pairs of genes in the same tissues across 30 different human tissues from the TiGER database of tissue-specific gene expression and regulation [75].

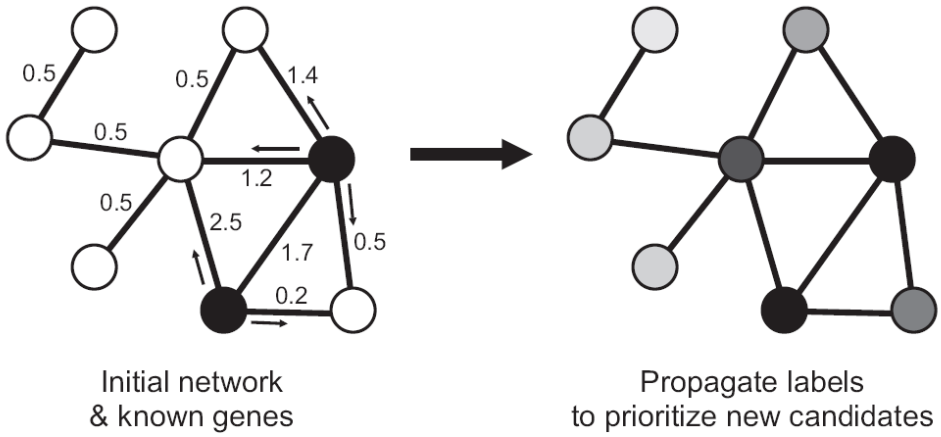


Figure 2.4: **A schematic figure of network-guided prioritization of candidate disease genes.** Given some known disease genes (black nodes), additional genes can be predicted by their (weighted) associations in the network, with more strongly connected genes being prioritized more highly (node shading).

pate in the same process as the known disease genes. We further considered four methods that diffuse disease associations across the network, considering both direct and indirect connections, similar to the methods considered in [77]. Two of these are mathematically related to Googles PageRank algorithm: (3) the Iterative Ranking method, in which a genes score is calculated from an initial score and the normalized scores of its neighbors, which, when updated over successive iterations, smear across the network linkages [74, 78], and (4) Gaussian field label propagation (Gaussian smoothing, for short), in which the difference between a genes initial and final scores and the weighted score difference between a gene and its neighbors are simultaneously minimized [73]. Finally, we considered (5) a clustering approach, using Markov clustering of genes based on simulation of stochastic

flow in the network [79], followed by ranking of each gene within a cluster for relevance by considering the sum of the genes edge-weights within the cluster relative to all of its edge- weights [74], and (6) a model based on electrical circuits [80], in which network edge weights are considered to be analogous to electrical conductance and disease implicated proteins are considered as ground nodes; candidate nodes are identified by modelling the application of current to the resulting circuit and measuring which nodes have the highest modelled current flow.

Figure 2.5 shows examples of ROC curves associating genes with several human diseases using the Iterative Ranking approach, showing high predictability for these cases. In order to systematically test if such predictability was common, and in order to judge the relative merits of the network diffusion approaches, we next evaluated a more comprehensive set of mouse phenotypes and human diseases.

We first evaluated the predictive power of HumanNet for genes associated (via orthology) with mouse mutational phenotypes, drawing upon the nearly 4,000 well annotated gene-phenotype associations identified in gene knock-out, gene trapping, and chemical mutagenesis experiments, and catalogued in the Mouse Genome Database (MGD) database [81]. In order to minimize the risk of circular predictions, we performed the tests using a version of the network lacking human literature-based linkages (i.e., no linkages by HS-CC or HS-LC). For each of the six approaches, we measured the network predictability for these mouse phenotypes using cross- validated ROC curve analysis, plotting the distributions of AUC (area under the ROC curve) scores for 3,374 gene sets associated with mouse phenotypes

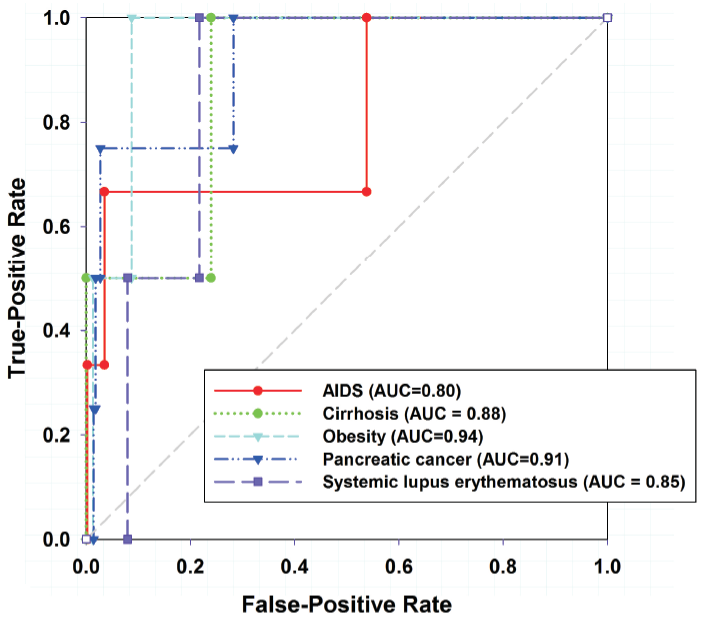


Figure 2.5: **Known genes associated with several human diseases are well predicted by the Iterative Ranking method for propagating disease labels across HumanNet, as measured using cross-validated ROC analysis.** The performance can be summarized as the area under the ROC curve (AUC), ranging from 0.5 (random) to 1.0 (perfect).

in Figure 2.6. HumanNet shows broad predictive ability of genes associated with specific mouse phenotypes, and is significantly better than expected by chance using each of the six algorithms. However, the closely related Gaussian smoothing and Iterative Ranking approaches perform comparably to each other, and significantly better than the other four approaches, indicating that there is a clear benefit to considering indirect connections as well as direct network connections.

Unlike mouse phenotypes, annotations for human disease genes are still extremely limited, spanning approximately 3,000 gene-disease linkages in human versus nearly 100,000 in mouse [11]. From annotations available at The Mendelian Inheritance in Man (OMIM) database, we selected 263 diseases with at least 3 associated genes. We tested the networks ability to associate genes with each of the 263 diseases using cross-validated ROC analysis, testing each of the six approaches, just as we did for mouse phenotypes (and again, using the version of the network lacking human literature-based linkages in order to avoid any potential circularity). We observed strong predictability for the human genetic diseases, with many disease gene sets predicted to high accuracy based upon gene-gene associations in the network (Figure 2.7). Again, the Iterative Ranking and Gaussian smoothing approaches performed similarly well, and significantly better than the other four approaches, confirming the general applicability of network label propagation for associating genes with human diseases and organism-level phenotypes.

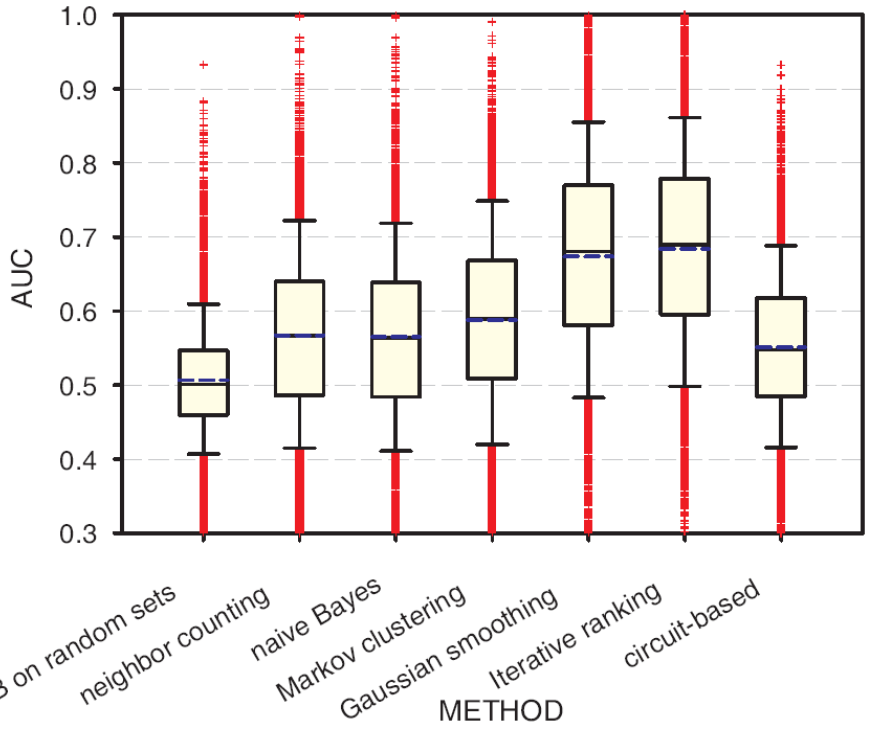


Figure 2.6: **Network GBA predictability of genes associated with 3,374 transgenic mouse phenotypes.** Bar-and-whiskers plots summarize the predictive performance (measured as cross-validated AUC) for each of six algorithms for using HumanNet to prioritize candidate disease genes. The Iterative Ranking and Gaussian smoothing approaches outperform the others by a significant margin, and show generally high predictability for more than three-quarters of the phenotypes tested. In bar-and-whiskers plots, the central horizontal line in the box indicates the median AUC and the boundaries of the box indicate the first and third quartiles of the AUC distribution, whiskers indicate the 10th and 90th percentiles, and plus signs indicate individual outliers. The mean AUC is plotted as a dashed blue horizontal line.

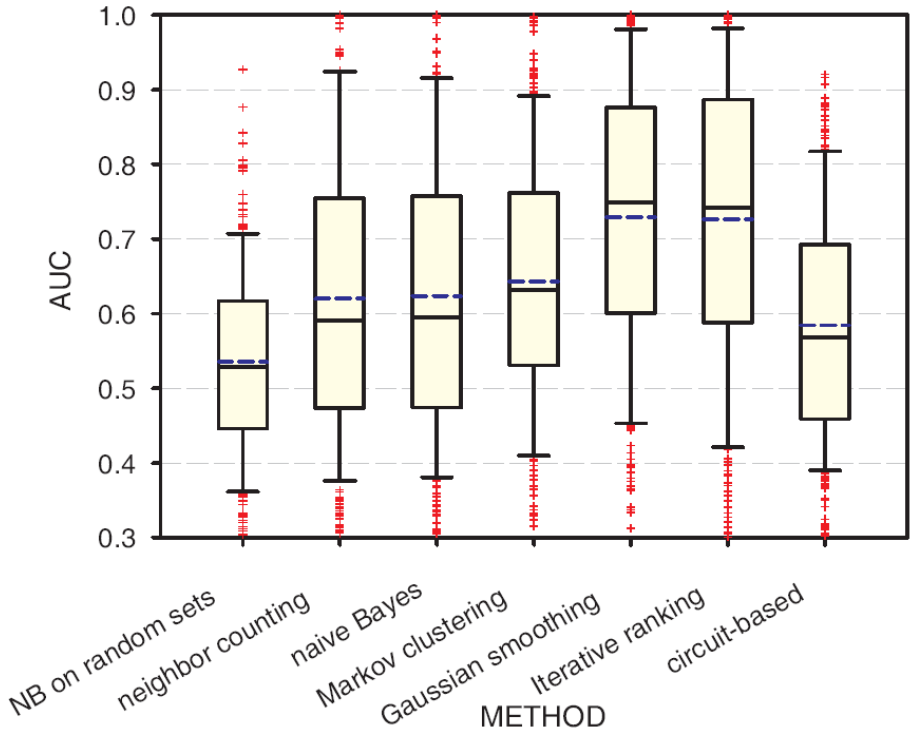


Figure 2.7: **Network GBA predictability of human diseases.** A related analysis to the one shown in Figure 2.6, but of human disease genes, assembled for 268 diseases from the OMIM database, shows similarly strong prediction strengths and the same relative ranking of algorithm performance.

Data from diverse sources is used to predict disease genes

We further investigated how the various data sets derived from high-throughput experiments and model organisms contribute to the mouse and human phenotype predictions. We examined predictions made by direct network connections using the naïve Bayes analysis and excluding the human literature-derived HS-LC and HS-CC datasets as for analysis in Figure 2.6 and 2.7. These contributions are visualized for the 20 most predictable mouse phenotypes and human diseases in Figure 2.8. Notably, datasets from worm and fly were strong contributors to the prediction of mouse phenotypes, as were data from human mRNA co-expression patterns (Figure 2.8A). Likewise, diverse datasets from yeast were strong contributors to a variety of well- predicted human diseases (Figure 2.8B). This demonstrates that most data sets contribute to the predictions, supporting the importance of data integration for effective disease gene identification.

Combining evidence from network guilt-by-association and genome-wide association studies

Given that network GBA is strongly predictive of human disease genes, a potentially powerful application of this approach is to combine the network GBA with the data from GWAS for direct discovery of human disease genes from patient and control populations. In order to use the information encoded by HumanNet, our method takes a slightly different approach from the SNP level tests used in the statistical analysis of GWAS today. Instead of focusing on single SNPs, we try to identify which genes and pathways might be involved in the disease. There are a number of reasons for this. First, even the SNPs that are identified in the traditional

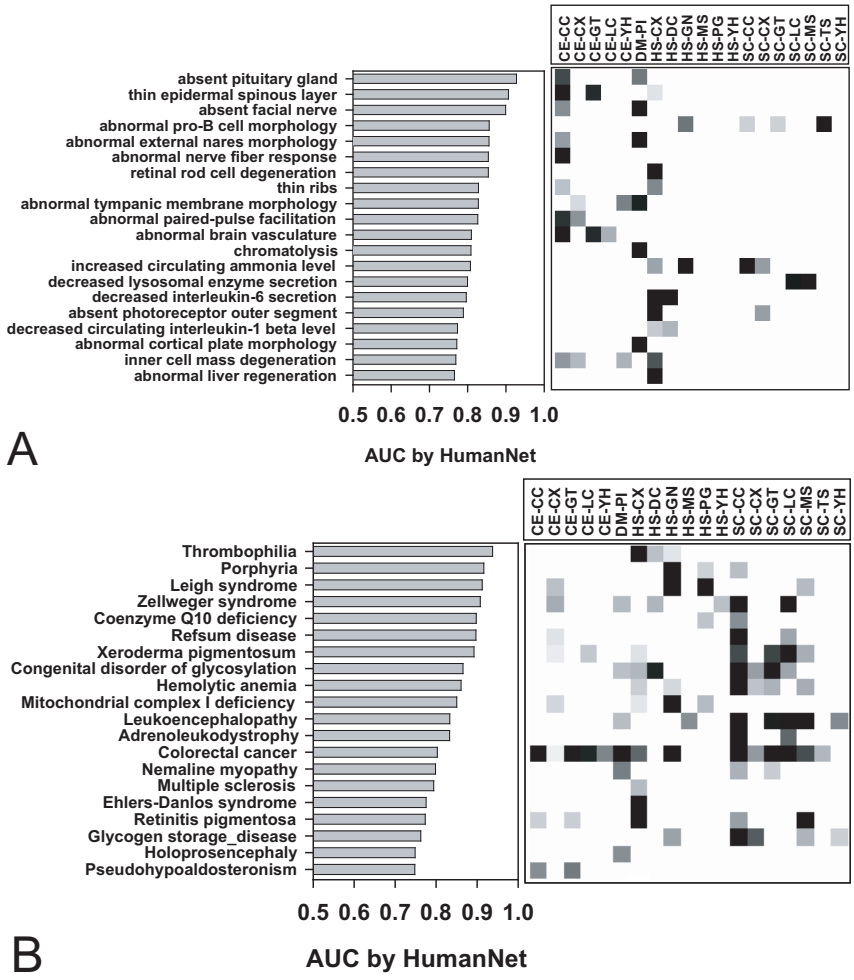


Figure 2.8: **The predictive power for loss-of-function phenotypes stems from a wide variety of data types integrated into HumanNet.** Prediction both of **(A)** genes associated with mouse phenotypes and **(B)** of genes associated with human diseases are supported by diverse lines of evidence, including, for example, fly and worm data contributing strongly to mouse phenotypes, and yeast data contributing to human diseases.

analysis are rarely thought to be the causal variants underlying the disease. This is due to the fact that the polymorphisms measured by GWAS have been chosen not for their biological significance, but for being the most informative of the surrounding region of the genome. Second, only a very small fraction of the genetic heredity of most diseases studied so far can be explained by the SNPs identified by GWAS [82]. This might be because a very large number of genes are involved in the diseases, or it might be because rarer variants cause a greater fraction of the heredity than previously thought [83, 84]. If it is due to the latter, we need to identify the regions of the genome where these rare mutations are located so that our search for such variants can be as efficient as possible. Our goal then is to identify genes and pathways of genes involved in the disease, not the marker SNPs most strongly correlated with the disease. Third, by taking a gene-centric approach, we can use the information encoded by HumanNet to improve our predictions. Finally, by working on the level of genes instead of SNPs, the method generalizes to future sequencing data, as long as the genetic variation can be associated with nearby genes.

If a GWAS finds a highly significant gene, it makes sense to attempt to identify the causal mechanism by which this gene influences the disease by looking at which pathways proteins encoded by this gene are active in. For example, this strategy leads to β-catenin expression and *WNT* signaling as a likely mechanism by which *TCF7L2* influences type 2 diabetes (for review, see Ref. [85]). By performing this type of pathway analysis automatically, it might also be possible to uncover genes that would not otherwise easily be found. This is especially true

for genes that fall just under the threshold of significance for the GWAS study, but which might be "rescued" by considering their interactions with the confident genes. Recent evidence for the case of human height shows that such minor contributions are common from polymorphisms falling below the significance threshold for association, but nonetheless contributing to total variation [31].

Unlike the GBA analyses considered above, for GWAS data, definite seed genes can rarely be found, particularly for the case where the only evidence for disease association comes from the GWAS itself. In order to make use of the information from the genes that are on the verge of being statistically significant, we implemented a "soft category assignment" for the GBA, where only genes that show a very strong signal are given full weight in the GBA. Notably, the performance of guilt-by-association in HumanNet is independent of the number of genes linked to the phenotype, which means that by varying the parameter that assigns weight in the GBA, we can include successively more genes that are increasingly less likely to truly be involved in the disease. We chose to base our method on the naïve Bayes GBA rather than the Iterative Ranking or Gaussian smoothing methods, since naïve Bayes gave superior recall in the highest precision regime, and the log odds output of the naïve Bayes can be combined with the log odds from the GWAS in a natural way.

Let S_i denote the total GBA score for a gene i, and denote by p_j the probability that some other gene j is involved in the disease. Suppose that j is connected in the functional network to i by a link of strength l_{ij}. It would then be natural to

assign a "soft" GBA contribution from gene j to gene i by

$$\Delta S_{i,j} = \left(p_j - (1 - p_j)\right) l_{ij},$$

which gives the total "soft" score S_i for gene i as

$$S_i = \sum_j \Delta S_{i,j} = \sum_j \left(2p_j - 1\right) l_{ij}.$$

However, this gives very poor results in practice, most likely because the network is only built on positive evidence. However, by only keeping positive contributions, we get good empirical results. Our "soft" GBA score is therefore

$$S_i = \sum_j \left(2p_j - 1\right) l_{ij},$$

where the sum is only over those j for which $2p_j > 1$.[1]

If we assume that the data from the GWAS and the data for the network are conditionally independent, we can then integrate them in a naïve Bayes framework, which, while not as strongly predictive as the iterative ranking and Gaussian smoothing strategies (Figure 2.4), was nonetheless quite robust for diverse diseases. The posterior log odds that gene i is involved in the disease are then

$$\ln O\left(i \in D | D_N D_{GWA}\right) = \sum \alpha_j l_{ij} + \ln O\left(i \in D | D_{GWA}\right),$$

where $\ln O\left(i \in D | D_{GWA}\right)$ is the log odds of association calculated from the GWAS data.

[1] Another natural way to take into account the fact that the network is built only on positive evidence would be to use p_j as a weight instead of $2p_j - 1$. In practice, however, this does not work well (data not shown).

Considering network linkages increases the power of genome-wide association studies

To evaluate whether the genes highlighted by this method actually are genes that are biologically relevant to diseases we used ROC analysis to compare how highly the combined GWAS/GBA method ranks the top candidates from meta-analyses for type 2 diabetes and Crohn's disease [86, 87], versus how highly those same genes are ranked by the Wellcome Trust study by itself [4]. These meta-analyses contain the Wellcome Trust data used for the predictions, but also incorporate data from a number of similar size studies, and have higher statistical power. For both type 2 diabetes and Crohn's, the Wellcome Trust study considered about 2,000 cases and 3,000 controls. For Crohn's, the meta-analysis considered 3,230 cases and 4,829 controls; for diabetes, 4,549 cases and 5,579 controls. To confirm that it really is the incorporation of the information encoded by the network that improves our predictions, we also compared these results with 200 randomly shuffled networks. As shown in Figures 2.9 and 2.11, the combined GWAS/GBA method clearly improves the ranking of the genes for both diseases, and does so over a wide range of parameter settings for the prior parameter.

Genes boosted in Crohn's

Prior to the Wellcome Trust study, strong association signals for Crohn's disease had been observed in *NOD2*, *IL23R*, *ATG16L1*, *ZNF365*, and in 5q31 and the gene desert 5p13.1. Furthermore, the Wellcome Trust study identified four more strong associations that were replicated in follow up studies. These were *IRGM*; a locus on chromosome 3 containing *BSN*, *MST1*, *MST1R*, *TRAIP* and some

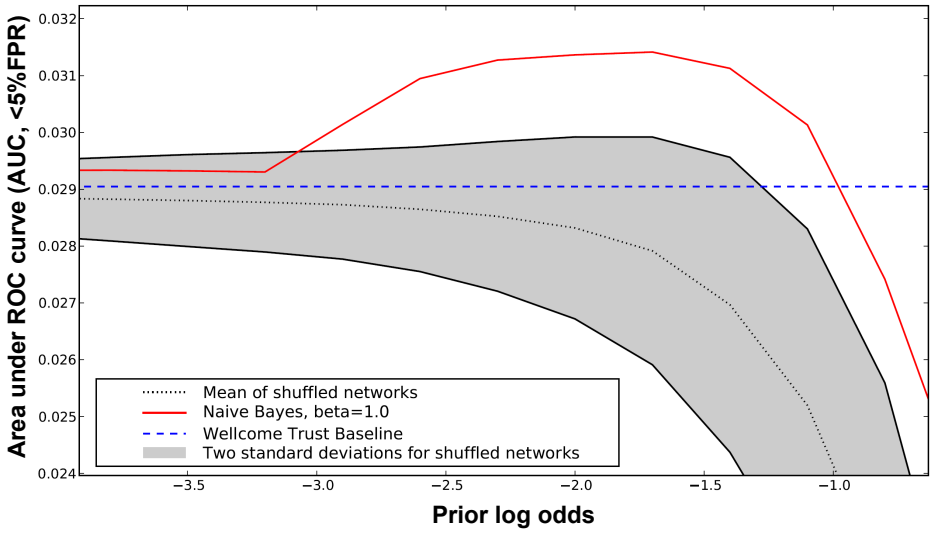


Figure 2.9: **Consideration of the human gene network boosts recovery of validated Crohns disease genes from GWAS analysis of 2000 cases and 3000 controls.** The performance improvement achieved by network-boosted GWAS relative to GWAS alone (Wellcome Trust Baseline, [4]), measuring performance as the area under a ROC curve up to 5%false positive rate (AUC, $<$ 5%FPR) for recovering the top 22 Crohns disease genes identified in a larger meta-analysis of 4549 cases and 5579 controls [86]. For the AUC ($<$ 5% FPR) measure of performance, a perfect predictor achieves a score of 0.05, while randompredictors score near 0.00125. The network boosted approach (colored red line) outperforms the GWAS alone (straight dashed blue line) over awide range of parameter values. For comparison we also show the results of network boosting when randomized networks are used, plotting the mean (dotted line) and range of performance (2 SD) for 1000 random trials.

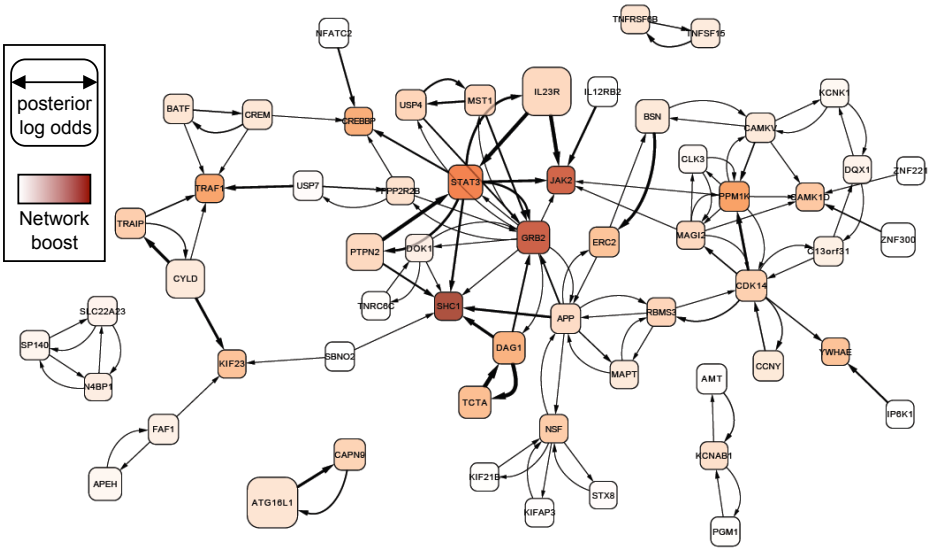


Figure 2.10: **Network of Crohn's disease candidate genes (rounded rectangles) identified from the combination of HumanNet and GWAS data,** visualized using Cytoscape [88]. The node size corresponds to the strength of the combined evidence from the Wellcome Trust Case Control Consortium (WTCCC) data and the network, and the intensity of the red color indicates how much the gene was boosted by the HumanNet GBA. HumanNet linkages are drawn as directed arrows connecting genes, with edge weight scaled by strength of boost contributed by the source to the sink. All genes are drawn with positive posterior log-odds when the prior log-odds of association are -1.7, except for network singletons, and the 50 highest scoring non-singleton genes are shown. Note the strong boost given to *GRB2* and *SHC1*, which are known to be involved in healing gastric ulcers (Pai *et al.* 1999), and to *JAK2* and *STAT3*, which were also identified in later meta-analyses (Van Limbergen *et al.* 2009).

other genes; *NKX2-3*, and finally, *PTPN2*. Moderate association was also seen in the regions 1q24, 5q23, 6p22, 6p21, 6q23, 7q36, 10p15, and 19q13, which contain a number of plausible candidate genes, such as *STAT3* and *TNFAIP3*.

Using the evaluation method described above, we saw a distinct increase in the top portion of the ROC curve for a wide range of values for the prior parameter centered at -1.7 (see Fig. 2.9). Using -1.7 as our value for the prior parameter, we then surveyed the gene groups that had strong network support. Interestingly, many of the gene clusters that emerged in this analysis showed strong connections with *TNF-alpha* signaling, which suggests multiple points of failure for the *TNF-alpha* pathway in Crohn's disease. We also note that one of the most successful drugs against Crohn's disease is the *TNF-alpha* antibody Infliximab.

IL23R, *STAT3*, *IL12RB2* and *JAK2* have all been indicated as candidate genes for Crohn's disease, probably affecting the disease through their involvement in the differentiation of Th17 cells [89]. These are strongly connected in our network, and therefore boost each other's rankings. For our choice of the free prior parameter, *STAT3* gets bumped from rank 17 to 7, and *JAK2* from rank 3139 to 89. Many of these are functionally connected in our network to both the adaptor protein *GRB2* (rank 99 to 12) and to its interaction partner *SHC1* (6125 to 63). *GRB2* and *SHC1* are also involved in gastric ulcer healing [90]. *GRB2* and *SHC1* are furthermore supported by their functional interactions with *PTPN2* and *MST1*, which probably affect Crohn's disease via their roles in the orchestration of the secondary immune response [89]. Lastly, *GRB2* is a binding partner to *TNFRI*, TNF receptor type I which can mediate a majority of TNF-alpha-dependent activities [91]. All of this

taken together indicates that *GRB2-SHC1* warrant further study as disease candidate genes for Crohn's disease.

The cluster containing *CYLD*, *TRAIP*, and *TRAF1* could also show a mechanism of action for Crohn's disease candidate genes. *CYLD* is located next to *NOD2* on chromosome 16. However, *Cyld-/-* knockout mice have an IBD phenotype [92], and *CYLD* has been shown to interact with *TRAIP* (*TRAF* interacting protein) by yeast two hybrid screens [93]. *TRAIP* is located in the 3p21 locus, which contains multiple independent signals for association with Crohn's disease [94]. Both of these genes are connected in HumanNet to *TRAF1*, TNF receptor-associated factor 1, which is involved in *TNF* signaling and *NF-kappaB* signaling.

We also see encouraging support of already known loci *TNFRSF6B* and *TNFSF15* are both known to be involved in Crohn's disease, and they are connected in HumanNet.

Another interesting gene association is given by *ATG16L1* and *CAPN9*, which boost each other. *ATG16L1* is involved in autophagy, and has been implicated in multiple GWAS. *CAPN9* is a stomach specific calpain, and mouse *Capn9-/-* knockouts are sensitive to gastric mucosal injury induced by ethanol administration [95]. This, together with their connection to *ATG16L1*, indicates that this is another plausible candidate gene for Crohn's disease.

Genes boosted in type 2 diabetes

Before the Wellcome Trust study, *PPARG*, *KCNJ11* and *TCF7L2* had all been identified as genes involved in type 2 diabetes through genome wide association

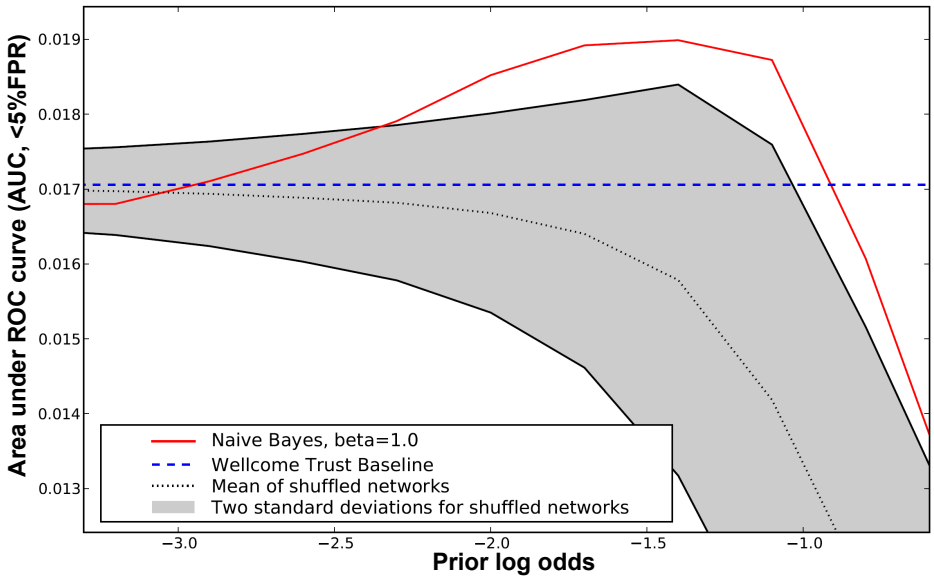


Figure 2.11: **Consideration of the human gene network boosts recovery of validated type 2 diabetes genes from GWAS analysis of 2000 patients and 3000 controls.** Plotted using the same conventions as in Figure 2.9, analyzing WTCCC GWAS data [4] for type 2 diabetes alone and in combination with HumanNet and measuring performance as AUC ($<$ 5% FPR) for recovering the top 20 genes from a type 2 diabetes meta-analysis of 4549 cases and 5579 controls [87]. As for Crohns disease, consideration of the network boosts performance across a wide range of parameter values.

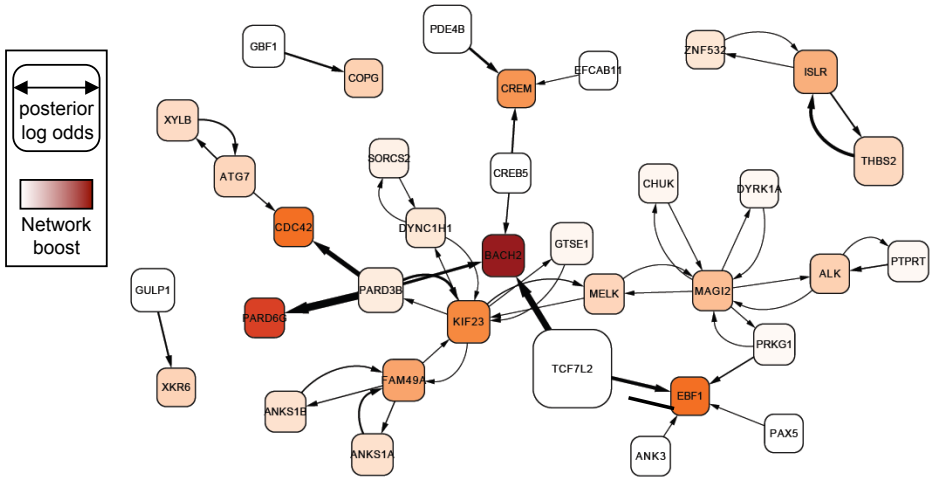


Figure 2.12: **Network of type 2 diabetes candidate genes (rounded rectangles) identified from the combination of HumanNet and GWAS data.** Consideration of the network strongly implicates the genes *CTNNB1* and *BACH2* in type 2 diabetes; *CTNNB1* is well studied in connection with type 2 diabetes and *BACH2* has been previously implicated in type 1 diabetes and celiac disease (e.g., [96, 97], but not type 2 diabetes.

studies and replicated in follow up studies (reviewed in [98]). The strongest candidate gene for type 2 diabetes, *TCF7L2*, was also the strongest signal seen in the Wellcome trust study, although the others were not so strong. However, the exact mechanism by which *TCF7L2* acts was not entirely clear. In our analysis, we find it directly connected to *BACH2*, a gene that has been repeatedly implicated in type 1 diabetes (e.g., [96]), but which has not yet been linked to type 2 diabetes. *BACH2* is among the genes most strongly boosted by network linkages, deriving additional signal from *CREB5* and *PARD3B*, which both score highly in the GWAS data. *PARD6G*, *PARD3B* and *CDC42* are also emphasized by the method. Notably, these genes form a complex with *PRKCZ* [99], a variant of which correlates with type 2 diabetes in Han Chinese [100]. *EBF1*, a known regulator of adipocyte differ-

entiation [101] is also strongly boosted by the network, supporting a possible role in type 2 diabetes.

Thus, for both Crohn's disease and type 2 diabetes, the combined GWAS / GBA approach both boosts genes that have support in other populations and that have been replicated in later meta-analyses, and highlights new connections between functionally connected genes among the genes that show moderate association to the disease.

2.3 Discussion

A new functional gene network for human genes

In order to test the general ability of a gene network to prioritize disease genes, particularly in conjunction with GWAS studies, we constructed a genome-scale functional network of human genes, incorporating diverse expression, protein interaction, genetic interaction, sequence, literature, and comparative genomics data, including both data collected directly from human genes as well as that from orthologous genes of yeast, worm, and fly. The resulting HumanNet gene network can be accessed through a web interface[2]. Using this interface, researchers can easily search the network using a set of seed genes of interest. The interface returns a list of genes ranked according to their connections to the seed genes, together with the evidence used to identify each coupling. The interactions and evidence can be downloaded, and a network visualization tool has been incorporated. All linkages can also be downloaded for independent analysis.

[2]Available at `http://www.functionalnet.org/humannet`.

Functional networks provide a general strategy for prioritizing disease genes

We demonstrate here that connectivity of human genes in an integrated functional network is a strong predictor of disease genes, both for cellular phenotypes and for diseases at the level of the whole organism. This predictability is strong even when considering only direct network connections, as shown both here and by related previous work (e.g., [9, 41, 42, 45]). We further show that algorithms developed originally for predicting gene function using gene networks also perform well at prioritizing candidate disease genes. Importantly, the consideration of indirect connections in diffusion algorithms, such as Iterative Ranking [78] and Gaussian smoothing [73]), greatly improves the correct identification of disease genes. Thus, knowing a few genes implicated in a disease, the networks offer a strong tool for prioritizing additional likely candidate genes.

One primary limitation of this approach is that genes must already be affiliated with the disease in order to predict new candidates. Typically, these seed genes would come from prior studies. However, we demonstrate that the approach is still valuable when used in combination with GWAS data, where no genes are definitively associated with the disease. Recent work also demonstrates that functional networks in worm and yeast can successfully predict genetic modifiers of genes [60] using the same network guilt-by-association approach. The effectiveness of this strategy in yeast and worms strongly supports using a human gene network in same manner to predict genes of synthetic or epistatic phenotypes. While relatively few such genetic interactions are known currently among human genes [102], as compared to the cases for yeast (e.g., [103–106]) or worm [107, 108], func-

tional gene networks offer a potential directed strategy for expanding current sets of human genetic interactions by prioritizing the tested interactions using gene networks, and our preliminary results demonstrating prediction of *KRAS* and *OCT4* modifiers (Figure 2.2B) support such an approach.

Tissue specificity profiles are shared by linked genes

One important characteristic of HumanNet is the tendency for linked genes to share specificity of expression in distinct tissues (Figure 2.3). The observation of tissue-specificity embedded in networks is consistent with our expectation for co-localization of proteins in the same functional modules (e.g., protein complexes and pathways) in specific cell types. However, this is nonetheless notable, since many of the raw datasets for network construction were not themselves tissue specific. For example, yeast-two-hybrid (Y2H) interactions are tested not in human cells but in yeast cells, and in fact, linkages derived only from Y2H do not show high tissue specificity (data not shown). Similarly, the phylogenetic profiling and gene neighbor comparative genomics approaches are strictly based on analysis of genome sequences and make no reference to tissue expression, nor do, for example, linkages inferred by homology from yeast. This trend for linked proteins in a genome-wide functional gene network to share tissue specificity has also been previously observed for worm and Arabidopsis gene networks [10, 46], and thus seems to be a result of the training process and integration of multiple data types correctly capturing the sorts of functional relationships reflected by the tissue specificity. A practical consequence is that a single genome-wide network of genes is

nonetheless able to successfully implicate genes in tissue- and cell-type specific disorders, as, for example, the case of liver cirrhosis genes, which are well predicted (AUC = 0.88; Figure 2.5).

Network-aided association studies: A general strategy for prioritizing genome-wide associations in human disease

The success of our approach suggests that analysis of GWAS data sets using gene networks offers a useful strategy for identifying both directly causal genes and even potential modifier loci in human disease, and since neither the pathway information encoded by the network nor the disease-association likelihoods that come out of the GWAS make any prior assumptions about the disease studied, this strategy is free from the study design bias that is inherent in candidate gene or candidate pathway analyses. The altered prioritization offered by the network-based association approach has the effect of shifting attention for follow-up studies to those genes (not SNPs) that are both best supported independently, and most likely to impinge upon the process(es) that are themselves best supported by the GWAS data, as determined from the current state of biological knowledge that has been objectively reconstructed and summarized in the gene network. Since this technique is gene focused and not SNP focused, it can be used with any future sequencing technology as long as the genetic variations can be associated with genes. In our analyses of Crohn's disease and type 2 diabetes, the network boosted identification of correct associations by 10% (measured in area under the first 5% of the ROC curve), which translated in practice to one to two genes more for these cases, a statistically significant, but not large effect. However, the organization of the

associated genes into processes offered a large practical benefit, such as focusing attention to *BACH2*, *CTNNB1*, and *EBF1*, which were not well-supported by the type 2 diabetes GWAS, but which were prominent network connectors between the well-supported genes. Furthermore, this boost is an effect of using the full network; individual sources of data do not provide nearly the same coverage and accuracy as the integrated network, and the kinds of data that is informative varies for the two different diseases studied (data not shown).

A second overall strategy also presents itself for integrating GBA and GWAS data sets, that of a candidate gene-based approach: It seems quite feasible to use GBA to known causal genes in order to select additional candidates, then to evaluate those candidate genes in a directed fashion, either by interrogating the GWAS data for associations involving these loci, or by directed sequencing of the candidate genes in patient populations. By focusing only on those genes ranked highly by GBA, the multiple testing explosion of typical GWAS is eased considerably, allowing for smaller patient samples to be tested and easier statistical significance thresholds to meet.

Concluding remarks

In summary, the approach outlined here provides a general method for prioritizing human disease genes, both for the case where seed genes associated with the disease are known already, and for the case where no such seed genes are known, but GWAS data for the disease is available. Our results suggest that the network will be useful for a considerable fraction of human diseases with genetic

components, and thus provides a general resource for diverse genetic diseases.

2.4 Methods

Construction of HumanNet

This study is based on 18,714 human Entrez genes with validated coding proteins (downloaded from NCBI; March 2007). Gene functional associations were trained using a reference set of gene pairs sharing Gene Ontology (GO) biological process annotations (downloaded from NCBI; March 2007). We used only annotations supported by experimental evidence: IDA (inferred from direct assay); IMP (inferred from mutant phenotype); IPI (inferred from protein interaction); IGI (inferred from genetic interaction); and TAS (traceable author statement). To minimize training bias, we excluded highly overrepresented annotations: (1) signal transduction (GO:0007165) (this term alone would otherwise account for 38% of total positive reference gene pairs); (2) three additional phosphorylation terms that have highly diverse biological roles, protein amino acid phosphorylation (GO:0006468), protein amino acid autophosphorylation (GO:0046777), and protein amino acid dephosphorylation (GO:0006470); and (3) all terms at the first and second levels of the GO hierarchy (assuming the term biological process is level zero). The resulting dataset of 270,704 reference gene pairs covers 5,369 (29%) human genes.

Functional associations were learned (as described in detail in the Supplemental Methods) in a supervised training framework using the log likelihood scoring (LLS) scheme of [10, 46], monitoring overtraining with 0.632 bootstrapping as in [10]. Gene associations from each separate dataset described below were optimized

to maximize performance as measured by precision-recall analysis, in accord with the rationales presented in [10, 46]. Multiple LLS for each gene pair were integrated using the weighted sum method with linearly decaying weights as in [10].

Analysis of tissue-specificity of network linkages

The similar tissue specificity of linked gene pairs was measured as the likelihood of co- occurrence of transcripts of pairs of genes in the same tissues, calculated as likelihood score

$$LLS = \ln\left(\frac{P(C|N)/P(\neg C|N)}{P(C|R)/P(\neg C|R)}\right),$$

where $P(C|N)$ and $P(\neg C|N)$ are probabilities that genes connected by the network (N) are co-expressed (C) and not co-expressed ($\neg C$) in the same tissue. $P(C|R)$ and $P(\neg C|R)$ represent similar calculations based on randomized networks (R), repeating calculations for 100 randomized networks. As a reference for tissue- specific expression, we collected 5,018 tissue-specific genes and their expression profiles across 30 different human tissues from the TiGER database of tissue-specific gene expression [75].

Implementation of network guilt-by-association algorithms

The naïve Bayes GBA algorithm was implemented as previously described [10]. Briefly, a gene score consists of the sum of LLSs to seed genes. For neighbor counting, the LLS sum is simplified to a count of neighboring seed genes. For Markov clustering, MCL software was downloaded from `www.micans.org/mcl` [109, 110]. We obtained network clusters using the default granularity settings.

The final score for a gene consists of the sum of the genes maximal coverage scores to clusters containing seeds. The coverage score is an MCL measure, comprised of the sum of edge weights from a node to a cluster, with larger edge weights rewarded. To obtain random scores for a phenotype set, we randomly selected from the genome a set of seeds of the same size, and performed naïve Bayes GBA as before.

The following methods were implemented in Matlab: GeneMANIA Gaussian field label propagation (Gaussian smoothing) was implemented as previously described [73]. Briefly, seeds were assigned initial scores of 1, and all others n/N, where n is the number of seeds and N is the total number of network genes. We then solved the system $y = (I + L)f$, where y is the set of initial scores, L is the graph Laplacian matrix of the network, and f is the set of final scores. The method for Iterative Ranking is derived in detail elsewhere [78]. However, rather than iteratively computing the final scores, we solved the system $y = (I - U)f$, where U is the matrix of network edges weighted by the sum of outgoing edges from each node. For the circuit based method, we followed the electrical model proposed previously [80]. Each edge in the network is treated as the conductance between the connecting nodes. The seed nodes are designated as the ground reference, and a current is simultaneously applied to all other nodes in the network. Using Kirchoff Laws, we solved for the voltage for each node. The final score for a node is the flow, or the nodes total current multiplied by its voltage.

Integrating the gene network with genome wide association study data

GWAS data came from the Wellcome Trust Case Control Consortium[4]. We selected the additive Bayes factor as a measure of association between SNPs and diseases, and represent each gene by the strongest association signal within 10 kb from the beginning or end of the gene. The same analysis for different cut-offs, varying from 0 to 250 kb, did not significantly change the boosting from the network.

We approximated the probability of a gene being involved in a disease by assuming that the space of possible hypotheses was limited to the null hypothesis and the additive hypothesis used for calculating the Bayes factors, and chose the value for the prior odds by optimizing the area under the first 5% of the area under the ROC curve. In general, we observed an improvement for prior ($\log_{10}$) odds ranging from roughly -2.5 to -1, corresponding to approximately 60 to 1,900 associated genes, respectively. Finally, in testing the effect of normalizing for node degree in the gene network, we observed a loss of performance, presumably because node degree does carry information for associating genes with diseases.

2.5 Acknowledgments

This work was supported by grants from the National Research Foundation of Korea (NRF), funded by the Korean government (MEST) (No. 2010-0017649), and POSCO TJ Park Science Fellowship to I.L and from the N.S.F., N.I.H., U.S. Army Research (58343-MA) and Welch (F1515) and Packard Foundations to E.M.M.. This study makes use of data generated by the Wellcome Trust Case-Control Con-

sortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

Chapter 3

Prediction of gene-disease associations using gene-phenotype associations in multiple distantly related species

This work is adapted from a manuscript we are preparing for submission, titled "Prediction of gene-disease associations using gene-phenotype associations in multiple distantly related species", for which John O. Woods and I are joint first authors.

In this paper we develop an extension to the initial phenolog idea that Edward Marcotte and Kris McGary that makes use of multiple phenotypes in many distant species. John Woods did some initial prototyping for the naïve Bayes scheme. I saw the parallell to recommender systems, designed the cross-validation scheme and prototyped the additive integration scheme. John worked out many of the problems orthology leads to, implemented the final version of the code, and researched the biological examples. John, Edward, and I wrote the paper together.

3.1 Background

Human traits, diseases, and phenotypes may have orthologous properties in other organisms, and such properties — typically phenotypes — are identifiable based on orthology of the underlying genes. Such orthologous phenotypes, or phe-

nologs, can be used to predict novel disease-causing genes; for example, McGary *et al.* identified *SEC23IP* as a neural crest effector, potentially involved in Waardenburg syndrome, based on the orthologous phenotype *negative gravitropism* in *Arabidopsis* [11].

Phenologs are a natural extension of the concept of deep homology: as a bird's wing and a human hand arose from a common ancestor structure with a common complement of genes and a similar developmental program[111], so also might less obviously related phenotypes derive from a common ancestor phenotype affiliated with an underlying conserved gene module. For example, certain mammalian neural crest defects and plant gravitropism share and partly arise from an ancient, highly conserved vesicle trafficking system.

We set out to improve upon the original phenologs algorithm, with a goal of ranking candidate genes relevant to specific traits and diseases, by developing an unsupervised method to search for similar phenotypes. We reasoned that gene – phenotype association predictions coming from multiply "nearby" (or high similarity) phenologs (e.g., 3.1), preferably across multiple species, should provide more predictive power than those from single phenologs. Candidate genes are ranked based on both the number and similarity of cognate phenotypes which involve those genes, in turn suggesting a prioritization for wet lab experiments.

We expanded upon the original phenolog study — which included gene – phenotype data from human, mouse, worm (*C. elegans*), baker's yeast, and *Arabidopsis thaliana* — by adding data from chicken, zebrafish, and even *E. coli*, as well as additional human and worm datasets. We show that phenotype data may come

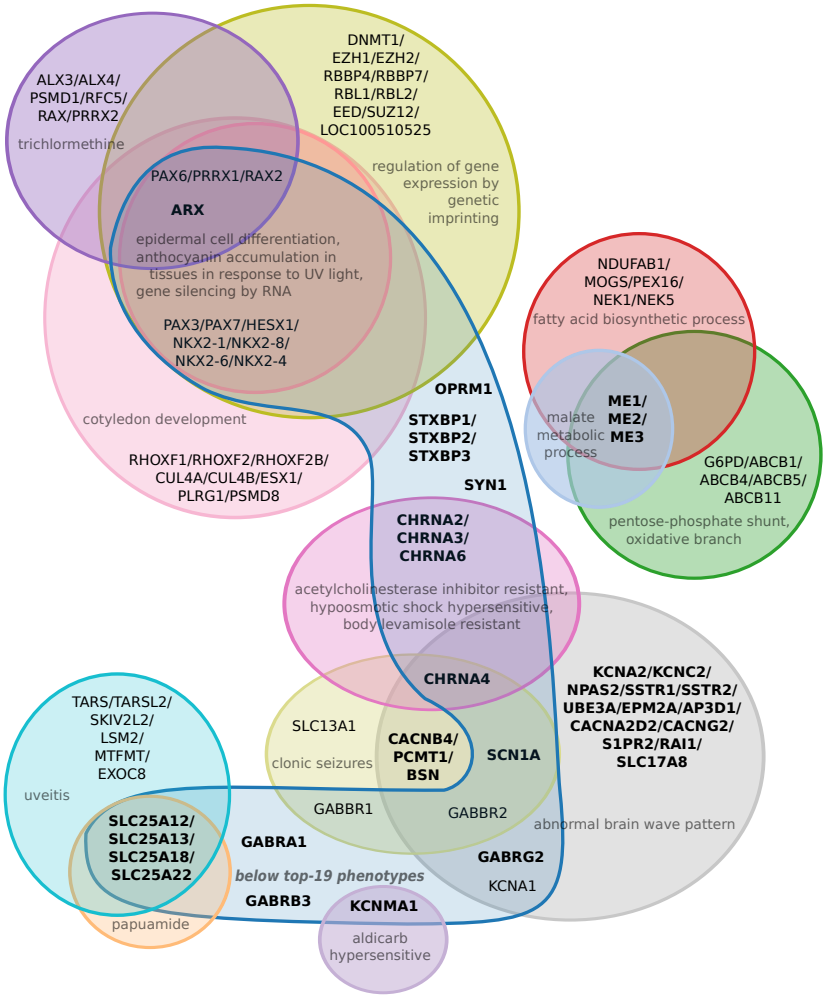


Figure 3.1: **A Venn diagram with predictions for epilepsy based on the 40 most similar phenotypes,** based on the Pearson sample correlation, and using cosine similarity as the weighting function. The nineteen closest phenotypes are each displayed separately, and the remaining twenty-one are aggregated into the category "below top-19 phenotypes." Paralogs are grouped together when they coincide at a prediction score. Genes in bold represent the orthogroups used in the search — that is, those groups of orthologous genes where one or more paralog was already associated with epilepsy in our database.

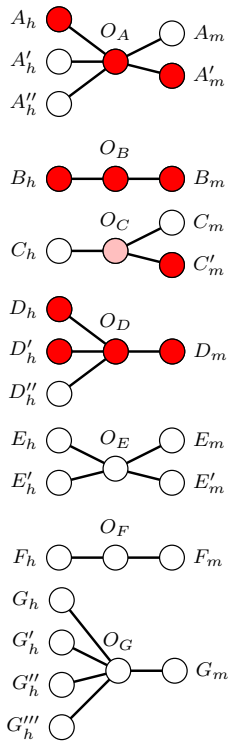
from a variety of sources, including BO biological processes and gene expression annotations, and that the integration of signal from multiply phenologs markedly improves the predictive power of the method.

3.2 Results

A key advantage to a neighborhood-based approach for predicting gene – phenotype associations is the ease with which non-obvious — and thus interesting — biological stories may be teased out. We demonstrate the process with epilepsy, a human syndrome; mouse *susceptibility to pharmacologically-induced seizures*, a related phenotype, using only *E. coli* data; and atrial fibrillation, the leading cause of arrhythmia in humans.

In addition to offering concrete predictions, we compare two classifiers for integrating phenologs (additive and naïve Bayes), across a variety of similarity or distance functions, and with different neighborhood (k) cutoffs. We also experiment with changing the weighting function used to assign prediction scores.

We experimented with two different frameworks for translating gene – phenotype associations between species, and designed a cross-validation scheme for each of these frameworks. In order to compare performance predicting phenotypes within a scheme, we used receiver-operating characteristic (ROC) curves and precision-recall plots.



(a) Orthogroups

	ϕ_h	ϕ_m
A_h	•	•
A'_h		•
A''_h		•
B_h	•	•
C_h		•
D_h	•	•
D'_h	•	•
D''_h		•
E_h		
E''_h		
F_h		
G_h		
G'_h		
G''_h		
G'''_h		
⋮		

(b) Human by gene

	ϕ_m	ϕ_h
A_m		•
A'_m	•	•
B_m	•	•
C_m	•	
C'_m		
D_m	•	•
E_m		
E'_m		
F_m		
G_m		
⋮		

(c) Mouse by gene

	ϕ_h	ϕ_m
A	•	•
B	•	•
C		•
D	•	•
E		
F		
G		

(d) By orthogroup

Figure 3.2: **The method for calculating phenolog overlaps** (or the similarity between two phenotypes coming from separate species) can bias the results, particularly for those species that are relatively isolated from the rest in the phylogenetic tree (e.g., plants). (a) Hypothetical orthogroups between some species h and another species m are pictured, considering a phenolog from that pair of species. The genes to the left of the orthogroup node are orthologous to the genes to the right, and vice-versa. Those genes indicated in red are associated with the phenolog. Matrices are displayed for predicting (b) species h using genes as rows, (c) species m using genes as rows, and (d) either species h or species m symmetrically using orthogroups as rows. In (b-d), bullets indicate an association between the gene and the phenotype ϕ_i. These matrix configurations produce radically different similarity scores using the hypergeometric CDF; only the orthogroup method produces scores that do not depend on the direction of the prediction.

Comparing genes between species

We tested two different frameworks for translating gene queries between species. For "gene-based" searching: given gene – phenotype interactions from a number of species, we encode the associations in matrices Φ_S, for S $\in$ {human, mouse, yeast, nematode, plant, zebrafish, fly, chicken}, where each Φ_S is defined by

$$(\Phi_S)_{ij} = \begin{cases} 1 & \text{if any ortholog in S of gene } i \\ & \text{is associated with phenotype } j, \\ 0 & \text{otherwise,} \end{cases}$$

where we used the INPARANOID algorithm [112] to approximate which genes in different organisms are orthologs of each other.

The INPARANOID algorithm discovers orthology relationships in the form of orthogroups (3.2A). For the method above we simply translate other species' gene – phenotype associations into the human gene – phenotype matrix by orthogroup, and compare the phenotype columns in terms of human genes (3.2). This scheme works well for closely-related species, where translations are 1:1 (e.g., a human gene has exactly one orthologous mouse gene).

However, when attempting to predict *Arabidopsis* phenotypes, we noticed that the translation process resulted in unusual scores, particularly when large orthogroups were involved (3.2). The genes-as-rows configuration also inflated performance, as measured by ROC plots, during cross-validation — primarily due to the high frequency with which plant gene expansions co-participate in a biological process.

Consequently, we began using an "orthogroup-based" matrix configura-

tion, which compared columns by observed orthogroups rather than by observed genes (3.2D),

$$(\Phi_S)_{ij} = \begin{cases} 1 & \text{if any gene in the orthogroup } i \\ & \text{is associated with phenotype } j, \\ 0 & \text{otherwise.} \end{cases}$$

The orthogroup-based matrix has the advantage of producing consistent, symmetric similarity scores irrespective of the direction of prediction; furthermore, these scores are not inflated by the presence of multiple phenotype observations in a single orthogroup.

Integration methods

Our goal is to construct a set of predictions for gene-phenotype associations X such that X_{ij} is higher for pairs where the gene is actually associated with the phenotype.

One way to incorporate information from multiple phenotypes is by measuring some kind of distance between pairs of phenotypes, and integrating the information from different phenotypes in such a way that "closer" phenotypes get more weight than phenotypes "far away". We tested two different ways of integrating this information — one multiplicative naïve Bayes scheme, and one additive method.

The naïve Bayes scheme we use was first described in the original phenolog paper [11], and can be written as follows:

$$X_{ij} = P(\text{gene } i \in \text{disease } j | k \text{ phenologs}) = 1 - \prod_{l=1}^{k} \left(1 - f_{ijl} w_{jl}\right) \tag{3.1}$$

where

$$f_{ijl} = P(\text{gene } i \in \text{disease } j | \text{phenotypes } j \text{ and } l \text{ are phenologs}) \tag{3.2}$$

$$w_{jl} = P(\text{phenotypes } j \text{ and } l \text{ are phenologs}) \tag{3.3}$$

As a proxy for the weighting function w_{jl} we have tried a wide range of measures that calculate a similarity or distance between two sets. Pearson sample correlation is a particularly popular option for expert recommendation systems. McGary *et al.* used the hypergeometric CDF, which gives the probability of seeing an overlap of size v or greater between phenotypes of size m and n, with N total orthogroups in the species pair.

For f_{ijl} we use v/n, the fraction of the number of genes common to both phenotypes j and l over the number of genes known to be involved in phenotype j, which empirically appears to be a good approximation for the probability that a candidate gene from a single phenolog will turn out to be a true positive.

We also developed an additive classifier. Whereas the naïve Bayes method multiplies distances or similarities as if they were probabilities, for the additive method we calculate X_{ij} for each gene-phenotype pair (i, j) by taking the sum over all phenotypes l, weighted by the similarity between j and the phenotype l, so

$$X_{ij} = \sum_k w_{jk} \Phi_{ik} = (\Phi w^T)_{ij}, \tag{3.4}$$

where Φ is a phenotype matrix and w is a weight matrix of phenotype – phenotype similarity scores.

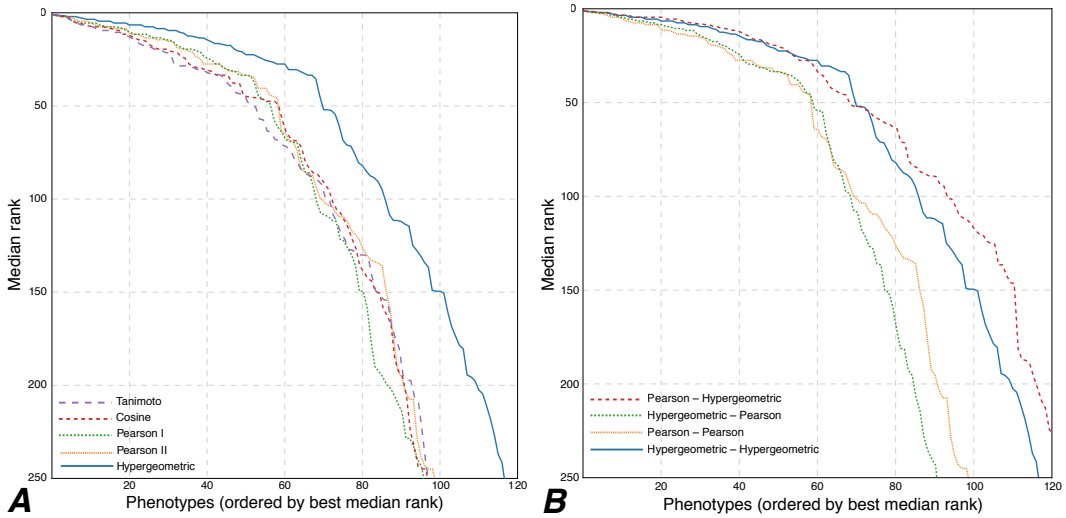


Figure 3.3: **Effect of distance measure choice for ordering and weighting.** Here we plot for how many diseases the median rank of the gene withheld during cross-validation stays at a certain level, using all available species, and integrating the results using the naïve Bayes scheme. In **A**, we vary the distance and weighting function (using the same measure for both). In **B**, we show the effect of varying the distance function independently from the weighting function. Here the first function in the legend is the distance function used for computing the k nearest neighbors, and the second is the weighting function w_{ij} from 3.1 and 3.4. As can be seen from the figure, a good distance function has more effect on performance than a good weighting function, but that the results can be improved slightly by using a combination: hypergeometric for distance, and Pearson for integration.

In addition to hypergeometric CDF and Pearson sample correlation, we tested Euclidean distance, taxicab (Manhattan) distance, cosine distance and Tanimoto coefficient as similarity measures, both for finding the k nearest neighbors and as weighting functions. Euclidean and Manhattan distance performed extremely poorly in five-fold cross-validation, so we excluded them from analyses in the orthogroup framework. Overall, Pearson and hypergeometric appear to have the most power for identifying nearby predictive phenologs (Figure 3.3A).

We also repeated the analysis while varying the distance function (used for searching) and holding the recommendation function (w) the same, and vice-versa (3.3B). Pearson sample correlation showed the best performance of the distance functions; however, we found that the hypergeometric CDF was the best weighting function for assigning prediction scores to genes.

We compared the naïve Bayes and additive classifiers, with the results shown in Figure 3.4. The performance in cross-validation is quite similar between the two classifiers, with the best version of the naïve Bayes classifier (using Pearson correlation for distance and hypergeometric for weighting) performing slightly better than the best additive one (using Pearson correlation and hypergeometric similarity). However, the additive classifier allows us to visualize and deconstruct the prediction into component phenotypes. We therefore chose to use the additive classifier for most predictions.

Varying the number of neighbors used (k) tended to affect lower-ordered predictions (e.g., the thousandth gene predicted for a disease) to a larger extent than top predictions. Figure 3.10 shows that even including the $k = 5$ nearest neighbors improves the results modestly — raising the number of diseases for which the withheld genes can be predicted at a top-100 median rank from around 50 to 80. Searching for the $k = 40$ nearest neighbors seems to offer no meaningful improvement over $k = 10$, unless one is interested in testing a thousand or so predictions; higher k values seem primarily to include withheld genes at lower median ranks.

Some phenotypes were absolutely unpredictable; however, several of these

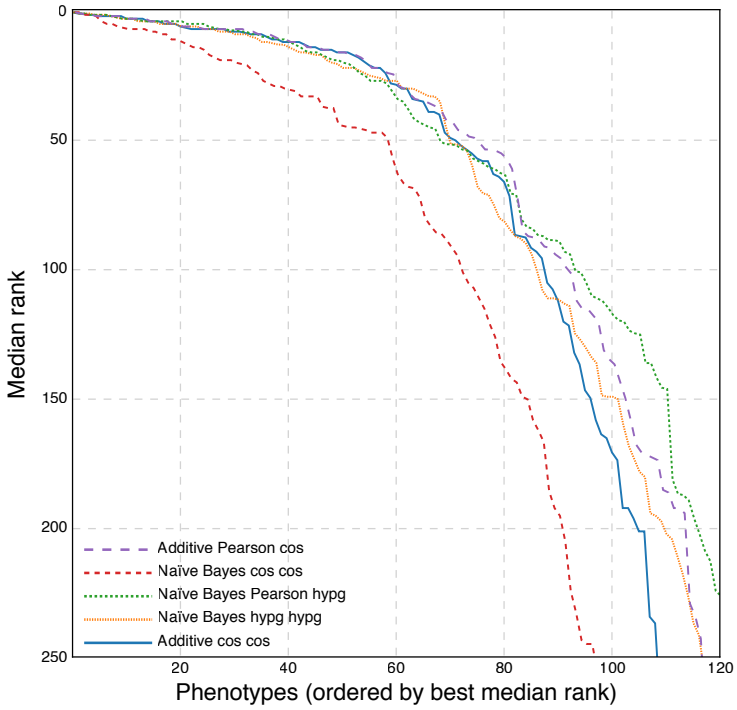


Figure 3.4: **Orthogroup-based matrix predictiveness.** Here we show a comparison of naïve Bayes and additive classifier predictions, which seem to have similar performance, using leave-one-out cross-validation. As in figure 3.3B, the first function in the legend is the distance function used for computing the k nearest neighbors, and the second is the weighting function w_{ij} from 3.1 and 3.4.

turned out to be combinations of diseases that were accidentally binned together in the initial version of our OMIM database (such as achromatopsia and achondroplasia). Blood type was one unpredictable — but properly binned — phenotype, likely due to the discrete nature of genes involved. In other words, each component of blood type (A and B, positive and negative) should be viewed as an independent monogenic phenotype.

The best similarity functions produced highly correlated predictions. Further, the best predictions of the worst classifiers were highly correlated with the best predictions of the top-performing classifiers. We thus concluded that the potential benefits of a fusion classifier would be modest at best, and difficult to measure at worst.

While similarity functions produced remarkably similar results, predictions coming from different species were much less strongly correlated (see 3.11). One potential improvement may be to weight phenotypes by species. While each species provides uncorrelated prediction information, the human disease predictions are dominated by mouse whenever that species is included — likely because of the highly correlated nature of the study of gene – phenotype associations in mouse and human.

Control: Randomized Matrices

To get an idea of how the scores are distributed when we attempt to predict from pure noise, we generated a series of random gene-based matrices. For each phenotype-column of cardinality p, we marked p randomly-drawn genes as

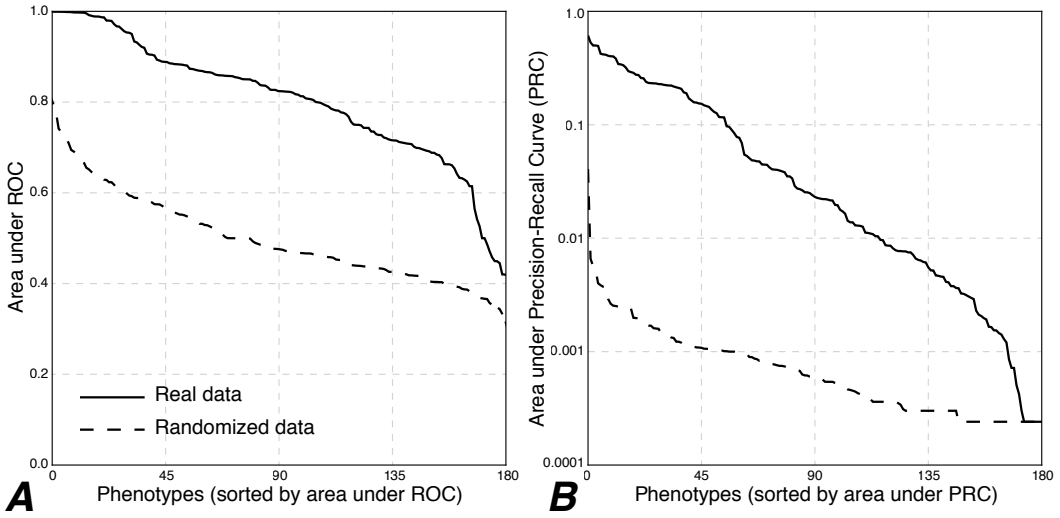


Figure 3.5: **Real vs randomized data.** Shown are ROC and precision-recall plots for $k = 100$ naïve Bayes using the hypergeometric weighting function, predicting human (OMIM) gene – disease associations from human, mouse, worm, fruit fly, yeast, and plant gene – phenotype association data. We restrict the evaluation to only those phenotypes with four or more known genes. The solid line shows the actual data, and the dashed line shows the result on similarly sized random gene sets.

observed. We attempted to predict phenotypes-of-interest from these randomized matrices using our regular classifiers (Figure 3.5 bottom).

As a random gene-based matrix would lack the structure which previously caused us to switch to orthogroup-based matrices — multiple paralogs participating in the same phenotype — we judged it unnecessary to randomize the orthogroup matrices.

Epilepsy

We chose epilepsy from our list of diseases because, despite offering ostensibly correct predictions, it actually scores somewhat poorly in cross-validation. In

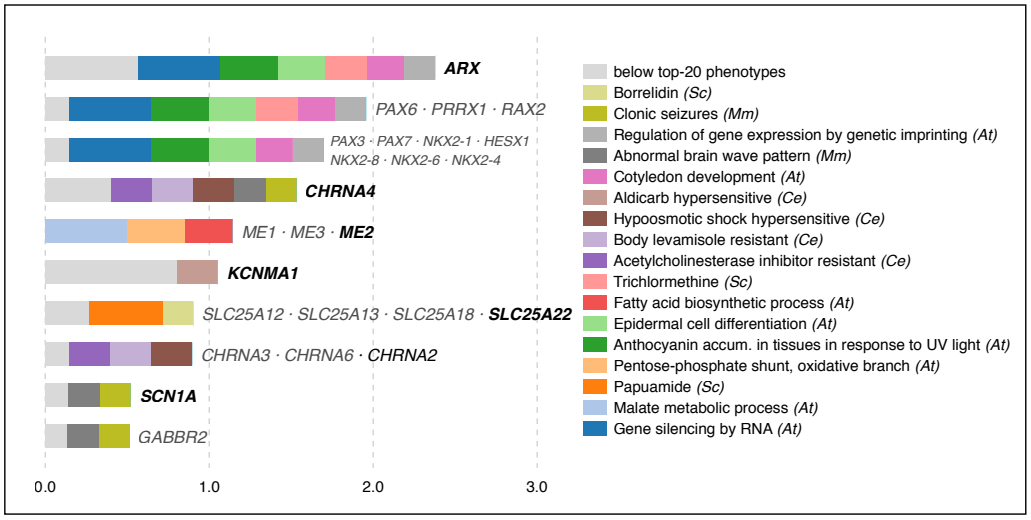


Figure 3.6: **Epilepsy** Each row of this chart represents a set of genes predicted with the same score. If a row's label (text) is printed in green, at least one of the genes predicted is already known to be involved; the other genes are likely to be paralogs. Rows with black-text labels are novel predictions. The depicted search makes predictions based on the $k = 40$ nearest neighbor phenotypes (from human, mouse, chicken, zebrafish, worm, yeast, and plant), and color codes the nineteen nearest neighbor phenotypes' contributions to each prediction (the remaining twenty-one are grouped in blue, as "below top-19 phenotypes"). The top scoring gene, *ARX*, is predicted primarily by Proud syndrome, hydranencephaly, and Partington's syndrome, all of which are human diseases characterized partially by seizures; but information is also drawn from a variety of plant phenotypes. These predictions were generated using an additive classifier for ease of visualization. The distance function is Pearson sample correlation, using cosine similarity as the weighting function w.

our initial three-fold leave-one-out test, only one of the three separately withheld genes was recovered at a reasonably testable rank (twelve, in this case).

Our method successfully identifies *GABBR1*, *GABBR2* [113], and *KCNA1* [114], which were absent from our database but known to be associated with the disorder. These were predicted primarily due to mouse phenotypes that resemble epilepsy (*clonic seizures* and *abnormal brain wave pattern*; see 3.1 and 3.6).

Top epilepsy predictions include *PAX6*, *PRRX1*, and *RAX2* (of which *PAX6* has been associated with seizures); and *PAX3*, *PAX7*, *HESX1*, and *NKX2-1*, *NKX2-4*,

NKX2-6, and *NKX2-8* (3.6)). Notably, *NKX2-1* is involved in mouse epilepsy [115], and *PAX3* appears in a region linked to the human version of the disease [116]); neither of these genes were in our database.

Interestingly, these predictions come from the *Arabidopsis* phenotypes *regulation of gene expression by genetic imprinting*, *cotyledon development*, *epidermal cell differentiation*, and *gene silencing by RNA*, as well as the yeast phenotype annotation for sensitivity to trichlormethine (nitrogen mustard, or *tris(2-chloroethyl)amine*).

To learn more about the general predictability of the epilepsy phenotype, we ran an expanded cross-validation, withholding each of the full set of 51 epilepsy genes in our database, and found that six genes could be predicted back — all within the top 120 ranks.

We wanted to know the extent to which predictions could be attributed to paralogy (shared orthogroup membership) with genes already associated in our database with epilepsy. *GABBR1* and *GABBR2* are each singleton orthogroup members, and are thus independently predicted. *KCNA1* and *KCNA2* emerged as paralogs following the human – worm divergence, but are predicted from non-worm phenotypes — and are therefore also independent predictions.

PAX6's plant – human paralogs make up the top three rank bins in 3.6. We suggest that even non-independent predictions are of use, provided they are accompanied by independent predictions — since, as mentioned, *PAX3*, *NKX2-1*, and *PAX6* are all associated to some degree with seizures and/or epilepsy. Indeed, the inclusion of species in which these genes are not paralogs offers additional resolution on predictions and demonstrates the utility of our method.

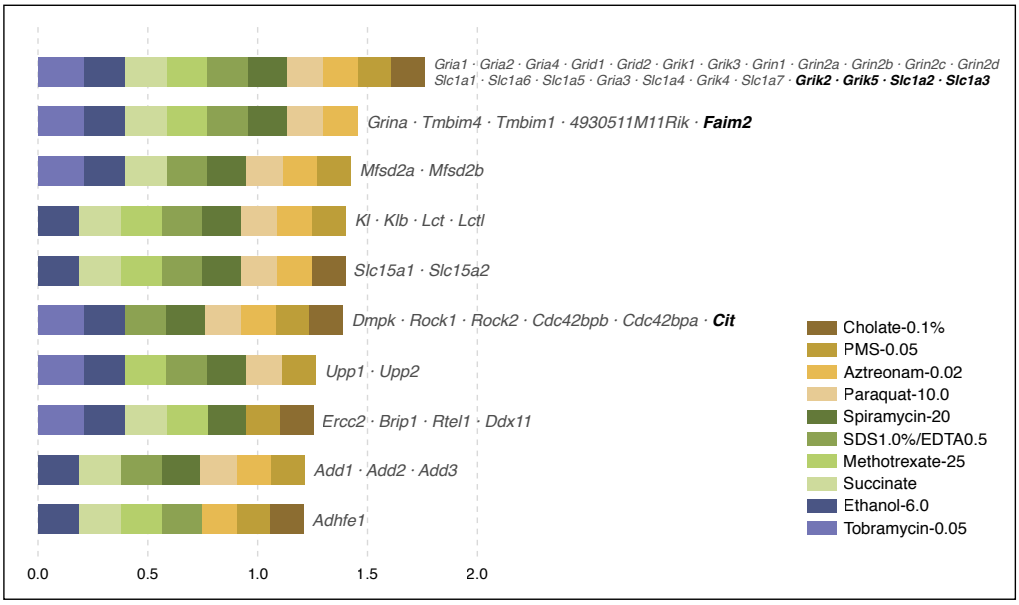


Figure 3.7: **Mouse seizures from *E. coli*** These mouse phenotype predictions are constructed from the $k = 10$ nearest neighbor *E. coli* phenotypes, using no other species. Predicting a eukaryote from a prokaryote is low-resolution, due firstly to evolutionary expansions of ancestral orthologs into larger orthogroups, and secondly to the tendency for some orthologs to vanish from certain species or become unrecognizable. Nevertheless, the probability of seeing an intersection of six or more orthogroups by chance, such as that between *tobramycin-0.05-unspecified* and the seizure phenotype, is 1.7×10^{-4} (without correction for multiple testing).

Predicting from *E. coli* — Pharmacologically-induced Seizures

We selected "pharmacologically-induced seizures" because this mouse phenotype could be predicted extraordinarily well from *E. coli* alone in cross-validation: eight of the forty-eight genes associated with this mouse phenotype could be predicted back when withheld.

These results are particularly impressive because they represent all six of the mouse – *E. coli* orthogroups associated with this seizure phenotype. Two of the

orthogroups (*Grik2*/*Grik5* and *Slc1a2*/*Slc1a3*) are in the top prediction ranking bin; additionally, *Faim2* is in the top hundred ranks (3.7).

The predicted gene *α-adducin*, likely the most promising result, is known to be reduced in the brains of rats experiencing kainate-induced seizures [117].

Another interesting prediction is *Sv2a* (*synaptic vesicle glycoprotein*). It was recently reported that a mutation in chicken *SV2A* leads to photosensitive reflex epilepsy [118]. Mouse *Sv2a* is a known binding site for levetiracetam, an antiepileptic drug [119], and $Sv2a^{-/-}$ mice experience seizures and die within three weeks of birth [120, 121].

We also searched for the source *E. coli* phenotypes (which happened to be chemicals) to see if these were associated with seizures. Standouts are ethanol — alcohol poisoning and alcohol withdrawal symptoms include seizures — as well as paraquat, which causes seizures and brain damage in rats [122], and aztreonam, which is a convulsant [123]. While nearly any compound is likely to lead to seizures if given in sufficient amounts, a back-of-the-envelope PubMed search for ten randomly chosen *E. coli* phenotypes from our database failed to turn up such clear associations.

Atrial Fibrillation

We selected atrial fibrillation (AF) for study because it had performed well in cross-validation in the gene – row configuration method. However, in the orthogroup-based cross-validation, only three of the eight genes associated could be predicted after being withheld. The removed genes were predicted at ranks 3–4, 15–16,

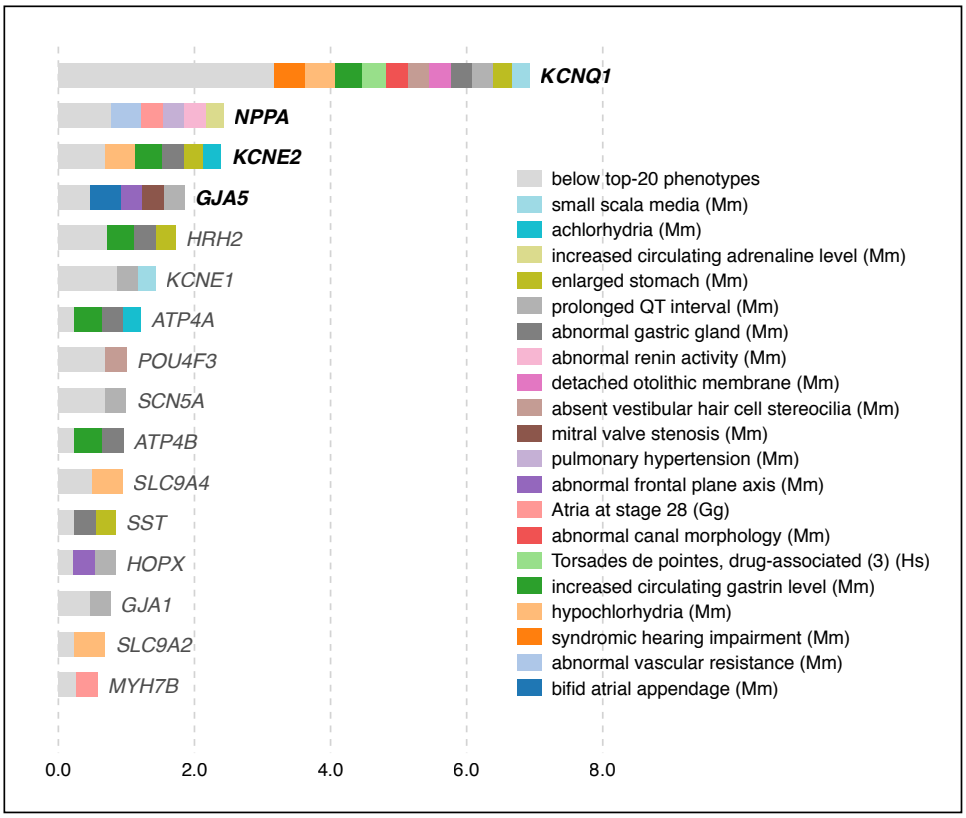


Figure 3.8: **Atrial fibrillation** These predictions are constructed in the same manner as those in 3.6. Limiting the search to $k = 40$ neighbors in this case means that all predictive phenotypes come from mouse and chicken, though other species were included in the analysis. Interestingly, few of the informative mouse and chicken phenotypes are related to the heart in any obvious manner.

and 81–94.

The top-ranked new prediction for atrial fibrillation (AF) is *histamine receptor H_2* (*HRH2*), largely contributed by gastrointestinal phenologs (3.8). Histamine has been known to act on heart cadence for over a hundred years [124]. However, an empirical link between heart and gastrointestinal function was established by the recent observation that histamine increases the heart rate in pythons during digestion [125] — regulation which occurs via the H_2 receptor [126–128], and which is apparently ubiquitous in vertebrates.

Similarly predicted are *ATP4A* and, further down the list, *ATP4B*, which are the α and β subunits of the H^+/K^+ ATPase. This proton pump is responsible for gastric acid secretion during digestion.

A somewhat speculative connection is offered by recent work, which showed cigarette smoke extracts cause an increase in the amount of H^+/K^+ ATPase in the stomach [129]. It is unclear — and worth testing — whether *ATP4A* and *ATP4B* are expressed in the heart. These genes could offer an additional route by which smoking contributes to heart problems.

Following *HRH2* and *ATP4A* is *HOPX*, or *homeodomain only protein x*, which is down-regulated during heart failure in humans [130]. It is not clear that HOPX is involved in AF per se, but worth exploring.

Next is *KCNE1*, based on orthologous phenotype *prolonged QT interval* — and seemingly also a factor in rare cases of atrial fibrillation [131–133].

GJA1 (*gap junction protein, α*1, also known as *connexin 43* or *Cx43*) is one of

the two most abundantly expressed connexins in the heart [134–136]. The other is *GJA5* (*connexin 40*), already associated in our database with AF. *Cx40* and *Cx43* seem to form heteromeric channels with different properties from homomeric channels [137]. *Cx43*, unlike *Cx40*, is essential for heart development and cardiac impulse conductance in mice [138]. Tuomi *et al.* observed that a dominant negative *Cx43* mutant causes severe AF [139]. Finally, atrial fibrillation was observed in a somatic mutation in human *GJA1* [140].

A similar story may be told for *SCN5A* (human cardiac sodium channel, voltage-gated, type V, α subunit), which is tied with GJA1. This sodium channel component has been associated with atrial fibrillation [141–143] but was missing from our database.

The top AF phenologs can be grouped into three basic categories: cardiac, gastric, and auditory. We have explored the first two categories, but have not considered genes from the third. We note that while Jervell and Lange-Nielsen syndrome (i.e., long QT syndrome) has been associated with deafness for half a century [144–147] via alleles of KCNQ1 [148] and KCNE1 [149], other genes may yet be involved [150]. Further, Belmont *et al.* write of "a growing appreciation for conditions that affect hearing and which are accompanied by significant cardiovascular disorders" [151].

Given the incredible success with which our method was able to predict atrial fibrillation genes — and with which it was able to identify potentially related disorders — exploration of additional candidates (e.g., *ATP4A/B*, *POU4F3*, and *S1PR2*) from 3.8 may be warranted.

Plant phenotypes — Response to Vernalization

Our search system is also capable of predicting gene – phenotype associations in species other than human. A number of factors reduce resolution in plant predictions. Firstly, while human phenotypes are predicted from other mammals and even other vertebrates — which are phylogenetically similar — there are no close neighbor species to *Arabidopsis* in our database.

Secondly, while 19,439 of the 28,002 human genes in our database have orthologs in other species, the ratio is less promising for *A. thaliana* phenolog predictions: 12,668 of 27,325 have orthologs. The cause is likely again the lack of other plants in our database, compared to the several vertebrates from which to draw information for *H. sapiens*.

Third and finally, the *Arabidopsis* genome contains a great deal of redundancy, as observed in [152]: 37.4% of proteins belong to families of more than five numbers, compared to 12.1% in fruit fly and 24.0% in worm. In orthology-based predictions, as with phenologs, there is no way to distinguish between such duplications — except perhaps by relying on paralogous phenotypes.

We determine phenotypes which may be predictable by cross-validating predictions produced from all non-plant species in the database (3.12). We make final predictions for *response to vernalization* (shown in Fig. 3.9).

We selected this phenotype because it scores better than most other plant phenotypes in cross-validation; seven of the fifteen genes in this plant phenotype can be predicted back at low rank when withheld, representing two or three or-

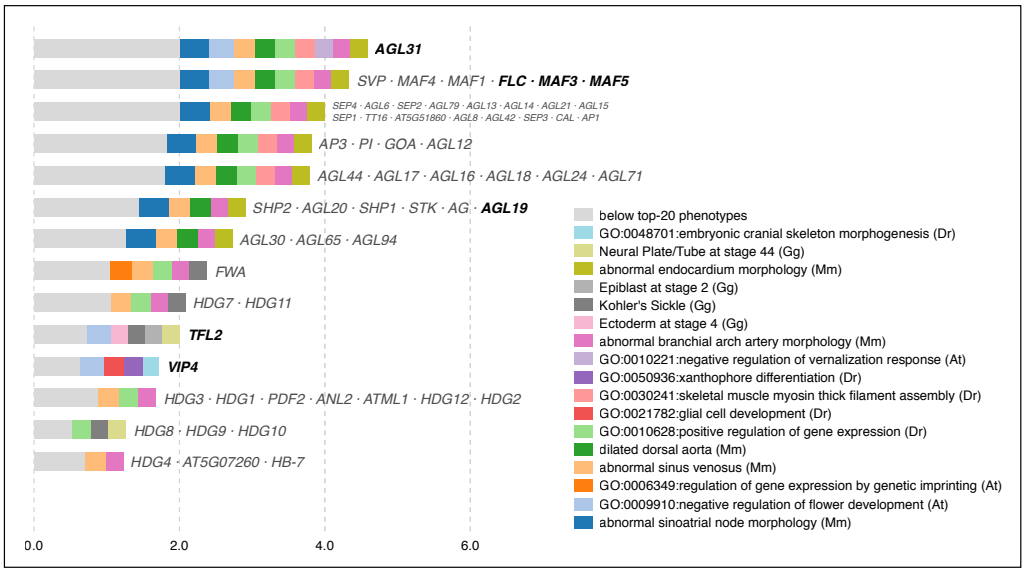


Figure 3.9: **Arabidopsis vernalization** Here, we demonstrate predictions for a plant phenotype, *response to vernalization*, while also demonstrating how including paralogous phenotypes may slightly enhance resolution. These predictions are drawn from phenotype data from each species in the database, with a neighborhood cutoff of $k = 40$. Due to the large gene expansions in plants, as well as the relatively large distance of *Arabidopsis* from other species in our database, paralogs are often ranked together. In the first two bins, a large gene expansion is split into separate ranks by information from an *Arabidopsis* phenotype (which is paralogous rather than orthologous). Those ranks labeled with green text include at least one previously known vernalization response gene (that is, a gene that was already linked with vernalization response in our database).

thogroups (about half of the total number of orthogroups) depending upon the source species considered.

Although we cannot cross-validate predictions from paralogous phenotypes, since they are not sufficiently independent, we feel that the inclusion of paralogous phenotype data improves the resolution at no perceivable cost.

Owing to the slow pace of study of plant genes, it is difficult to determine whether predicted genes (3.9) are correctly implicated.

However, we note at least two interesting predictions. One of these, *EMF2* — which appears to be associated with vernalization-mediated flowering by its interaction with *CLF* [153] — is predicted based on seemingly unrelated orthologous mouse and human phenotypes (abnormal chorion morphology and endometrial cancer, respectively).

EMF2 is paralogous with known vernalization gene *VRN2*; however, it is ranked ahead of *VRN2* by its association with the related plant phenotype *negative regulation of flower development*. That *EMF2* was boosted by a potential paralogous phenotype supports the hypothesis that paralogous phenotypes are similarly useful to orthologous phenotypes in predicting gene function.

The second interesting prediction is *FWA* and its several paralogs (*HDG1–4*, *HDG7–12*, *PDF2*, *ANL2*, *ATML1*, *AT5G07260*, and *HB-7*).

Certain *FWA* mutants produce a vernalization-insensitivity phenotype [154, 155]. Candidates *ANL2* and *PDF2* both have late flowering phenotypes [156, 157] markedly similar to that of *FWA* [158]. That discovery lends additional support for

paralogous phenotypes, as neither *FWA* nor *PDF2* were associated in our database with regulation of flower development — but our method successfully identified *negative regulation of flower development* as a potential phenolog.

Fruit Fly Phenotypes

While FlyBase holds a wealth of knowledge on gene – phenotype associations, we found it difficult to leverage for our goals. The controlled vocabulary is designed for searching for classes of phenotypes rather than the phenotypes themselves.

Unfortunately, the only way to connect a phenotypic class annotation to an anatomical location or developmental stage is by allele and literature reference — if these are given at all. While it was possible to predict some human diseases based on fruit fly phenotypes from FlyBase, the results were difficult to interpret.

Given the resulting noisiness and sheer quantity of data, we anticipated that correct predictions would be attributable to multiple hypothesis testing, and chose to exclude fruit fly results.

Pharmogenomics Knowledge Base Phenotypes

The Pharmacogenomics Knowledge Base, PharmGKB, offers a wealth of gene – phenotype associations. However, these associations are typically made by way of a drug used to treat a disease — which may act on a target which is not directly involved in the disease.

While we chose not to include PharmGKB results, the data is available

alongside the source code.

3.3 Conclusion

We set out to improve upon the results of the original phenolog project by unifying information from a "neighborhood" of phenotypes surrounding the desired disease. Our method produces ranked predictions for a large percentage of human diseases in OMIM, as well as for plant biological process-based phenotypes.

Furthermore, we were able to demonstrate the correct prediction of at least one gene associated with the mouse phenotype *pharmacologically-induced seizures* using only phenologs from *E. coli*. While McGary *et al.* demonstrated the existence of deep homology between mice and single-celled eukaryotes, our work suggests that examples of deep homology exist — and may even offer useful predictions — between prokaryotes and eukaryotes.

We also demonstrate that the term "phenotype" may be interpreted broadly when incorporating gene-association data for phenolog-based predictions. Gene Ontology biological processes are one potential source; another is annotations for *in situ* hybridization experiments, such as GEISHA.

We give a number of concrete gene predictions for the human diseases atrial fibrillation and epilepsy, and show how phenologs may be used to generate hypotheses and a biological context that correctly connect categories of diseases, such as disorders of the heart, stomach, and sensorineural system.

3.4 Methods

Cross-validation

For the gene-based matrix, we compared classifiers and metrics using *n*-fold cross-validation, and calculated receiver-operating characteristic (ROC) and precision–recall curves for each disease or phenotype to be predicted. Classifiers could be represented by arrays of area-under-the-curve measurements.

With the orthogroup-based matrix we chose a simpler and faster "leave one out" cross-validation scheme, where one observed gene association was hidden for each disease. Noting that some orthogroups have multiple genes associated with the same phenotype, we also hid any orthogroups associated with hidden genes. Since a gene may be part of one orthogroup for each species included in the search, we measured the rank of predicted genes rather than predicted orthogroups. When multiple genes were predicted with the same score, the mean rank was used.

The leave-one-out procedure was repeated three times for each phenotype, taking the median hidden gene rank to be representative of the classifier – phenotype performance.

Additional phenotype data

In addition to those databases described in [11], we incorporated orthology and gene – phenotype data from a variety of additional species.

New *C. elegans* phenotypes came from Green *et al.* [159] and were broken down into two datasets: *green-broad* and *green-specific*.

We added, without modification or filtering, a second human dataset (*pharmgkb*) from the Pharmacogenomics Knowledge Base [160].

E. coli phenotypes were taken on May 20, 2011, from the file 'coli_FinalData2.txt' [161]. Each gene's phenomic profile was sorted by score, assigning both the top and bottom forty conditions to the gene. Thus, each condition was considered to be a phenotype, and the genes associated with that phenotype were those genes whose growth was most affected — either positively or negatively — in the corresponding condition.

Fruit fly phenotypes came from FlyBase [162]. We attempted to match anatomical annotations for mutant phenotypes to annotations from the *phenotypic class* ontology, joining on allele and publication.

Zebrafish phenotypes consisted of gene ontology (GO) biological processes from ZFIN [163], keeping only those annotations with evidence types of *IMP*, *IDA*, *IPI*, *IGI*, *TAS*, *NAS*, *IC*, and *IEP* — the same procedure used for *Arabidopsis* phenotypes, obtained from TAIR [164]. These evidence types were selected so as to avoid the inclusion of annotations that originated directly from knowledge of other model organisms.

For chicken (*Gallus gallus*) phenotypes, we utilized *in situ* hybridization annotations from GEISHA [165], kindly provided in XML format on June 24, 2011. If there were more than fifty genes associated with a specific location and more than three at a specific state at that location, a new phenotype was created ("*anatomical location* at stage *x*"); and regardless, each location became an independent phenotype.

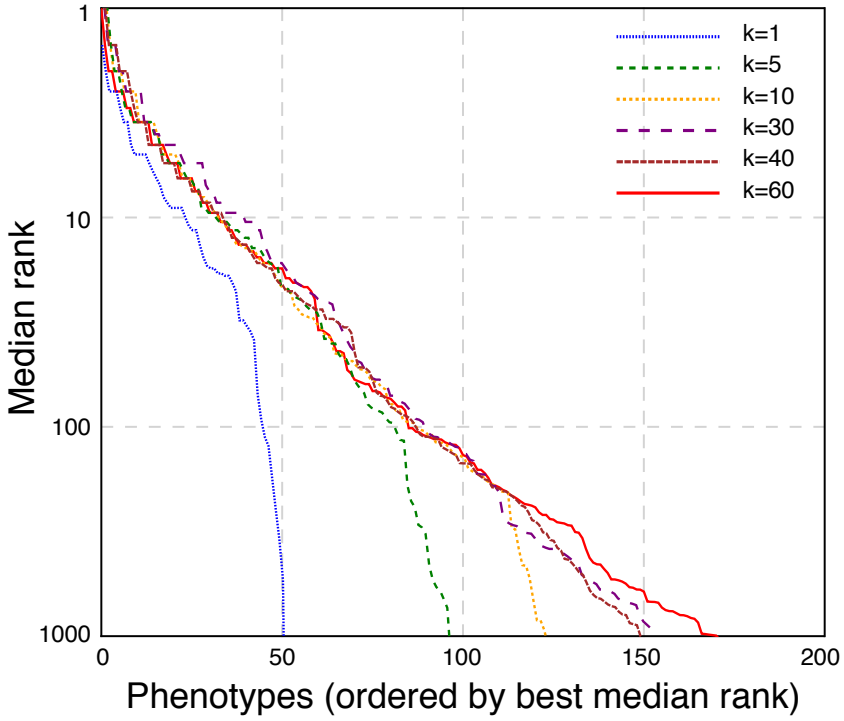


Figure 3.10: **Effect of *k* on Predictiveness.** We compare different *k*-values in the neighborhood search for phenologs. This graphic demonstrates the effect (and diminishing returns) of increasing *k* on recovery of withheld genes.

We defined phenotypes as gene – expression associations in specific anatomical locations. For those locations with more than fifty genes annotated, we created additional phenotypes for each stage with greater than three associated genes.

While developing the orthogroup-based matrix configuration, we updated our human OMIM dataset. Thus, predictions from that implementation utilized a somewhat different database than those used by McGary et al.

We excluded any phenotypes with fewer than three associated genes.

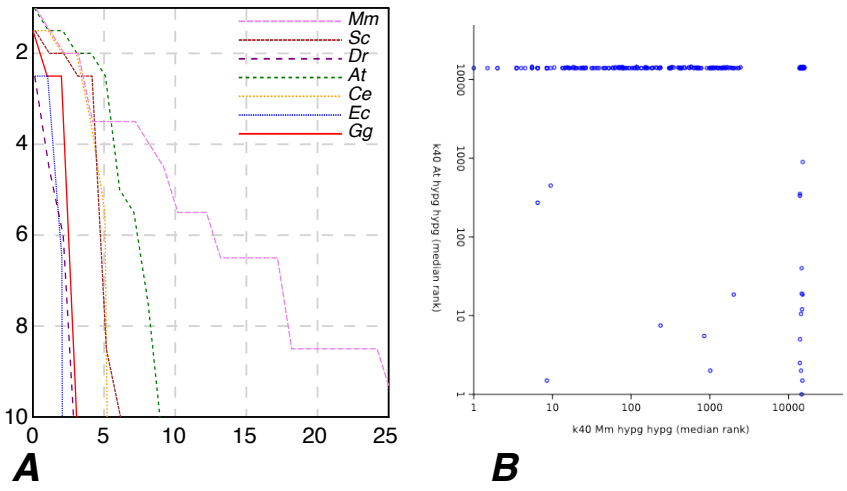


Figure 3.11: **Contributions by Individual Species.** ***A*** demonstrates that each phenotype offers some sort of information for prediction of human disease genes; mouse data seem to offer the most information about human diseases, as one would expect from the quality of the data and the proximity of the species in the phylogenetic tree. *Arabidopsis*, which is the furthest species from human in our database, unexpectedly provides as much information as mouse on top predictions, and is second at higher ranks. ***B*** A scatter plot qhich demonstrates that the information offered by each species (in this case mouse and *Arabidopsis* is highly independent, and suggests that integrating data from multiple species may be useful.

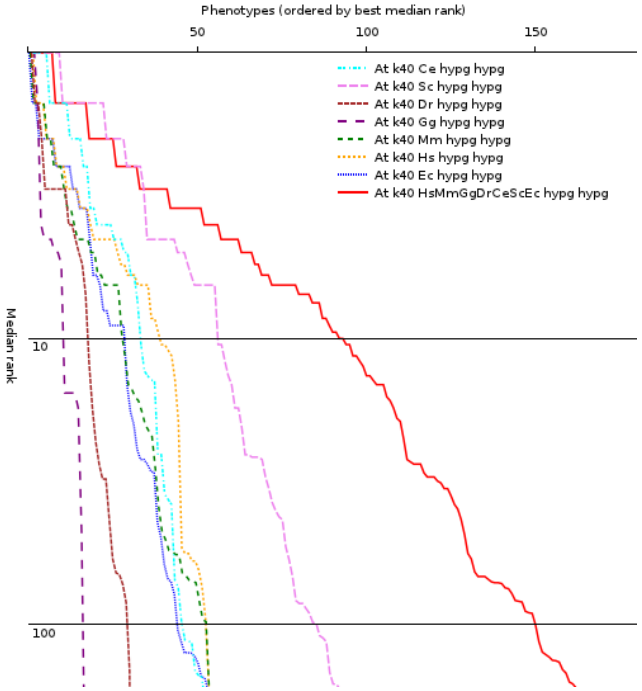


Figure 3.12: **Predicting Plant Phenotypes.** This figure mirrors 3.11A, but demonstrates the prediction of *Arabidopsis* phenotypes from individual species (rather than human diseases from individual species). The red solid line shows the combined performance of predictions using all species except *Arabidopsis*. Yeast appears to be the most useful individual species for predicting plant phenotypes.

Chapter 4

Prediction and validation of gene-disease associations using network methods

This part grew out of collaborative filtering inspired work in the previous chapter, and was done in collaboration with Nagarajan Natarajan in Inderjit Dhillon's lab, and Ambuj Tewari (formerly in Dhillon's lab, now a professor at University of Michigan in Ann Arbor).

Nagarajan and I have worked very closely on this project, and bounced ideas back and forth. The idea for the support vector machine formalism came from Naga and Ambuj, while I have done the analysis of the results.

4.1 Abstract

Correctly identifying associations of genes and diseases has long been a goal in biology. With the emergence of large-scale gene-phenotype association datasets in biology, we can leverage statistical and machine learning methods to help us achieve this goal. In this paper, we present two methods for predicting gene-disease associations based on functional gene associations and gene-phenotype associations in model organisms. The first method, Katz, is motivated from its success in social network link prediction, and is very closely related to some of the recent methods proposed for gene-disease association inference. The second method,

called CATAPULT(Combining dATa Across species using Positive-Unlabeled Learning Techniques), is a supervised machine learning method that uses a *biased* support vector machine where the features are derived from walks in a *heterogeneous* gene-trait network. We study the performance of the proposed methods and related state-of-the-art methods using two different evaluation strategies, on two distinct data sets, namely OMIM phenotypes and drug-target interactions. We also present qualitative analysis of predictions made by the methods and demonstrate that using one of our two evaluation strategies based on *singleton* traits yields performance comparisons that match how the predictions made by the respective methods compare qualitatively.

4.2 Introduction

Correctly predicting new gene-disease associations has long been an important goal in computational biology. One very successful strategy has been the so-called guilt-by-association (GBA) approach, in which new candidate genes are found through their association with genes already known to be involved in the condition studied. This association can in practice be derived from many different types of data. Goh *et al.*[166] construct a network where genes are connected if they are associated with the same disease, whereas Tian *et al.*[167] combine protein interactions, genetic interactions, and gene expression correlation, and Ulitsky and Shamir[168] combine interactions from published networks and yeast two-hybrid experiments.

One of the most commonly used kinds of association is derived from di-

rect protein-protein interactions, such as the ones curated by the Human Reference Protein Database (HPRD) [169]. The last few years have seen a number of methods that have extended the association from just direct protein interactions to more distant connections in various ways, such as CIPHER [170], GeneWalker [171], Prince [172] and RWRH [173]. One kind of network that has proven to be particularly useful for predicting biological function is the functional interaction network, where a pair of genes is connected based on the integrated evidence from a wide array of information sources, as seen in Lee at al.[8]. These have been used to associate genes with phenotypes in model organisms [9,73] and in humans [44,45]. A recently published network, HumanNet, has been used to refine predictions from genome-wide association studies [21]. Since functional gene interaction networks aggregate many different types of information, they can achieve much greater coverage than pure protein-protein interaction networks.

Alternatively, we can also think of the gene-disease association problem as a *supervised learning problem*, where each gene-disease pair is represented by a number of derived features (explicitly or implicitly using a Kernel function) and then a classifier is learnt to distinguish "positive" associations from "negatives", using previously studied gene-disease associations, and *unknown* gene-disease pairs as training data. Such an approach is taken by the recent ProDiGe method [174], which integrates a wide variety of heterogeneous data sets and uses support vector machines (SVMs) to identify potential gene-disease associations.

In the past decades, the growth of gene-phenotype associations in model species has been explosive, which suggests an alternative way to find candidate

genes for human diseases. McGary *et al.*[11] used this treasure trove of information to find surprising connections between model species phenotypes and human diseases by looking for pairs of human diseases and model phenotypes that share a higher than expected number of orthologous genes. In this way, a number of new, and often surprising, model systems were found for human diseases. For instance, the human neural crest related developmental disorder Waardenburg syndrome shares gene modules with gravitropism (the ability to detect up and down) in plants, and mammalian angiogenesis has been found to involve the same pathways as lovastatin sensitivity in yeast. This model species information represents yet another form of functional connection that can be used for gene-phenotype association.

In this work, we first propose two distinct but related GBA methods. One is based on the Katz method[175] that has been successfully applied for link prediction in social networks. The method is based on integrating functional gene interaction networks with model species phenotype data and computing a measure of similarity based on walks of different lengths between gene and phenotype node pairs. The second method, which we call CATAPULT (Combining dATa Across species using Positive-Unlabeled Learning Techniques) is a supervised learning method, wherein we represent gene-phenotype pairs in a feature space derived from *hybrid* walks through the heterogeneous network used by Katz. The supervised learning method falls under a class of learning methods called *Positive-Unlabeled* learning methods (ProDiGe [174] also belongs in this class) since the learning task has only positive and *unlabeled* examples (and *no* negative examples).

The method naturally generalizes the computation of Katz on a heterogeneous network by learning appropriate feature weights.

To determine if a computational method truly associates genes with diseases, biological validation of the predicted associations – often by knockout studies in model systems, or through sequencing of patients – is needed. Since these can be expensive and hard to do in a high throughput way, it is common to measure the performance of GBA methods through cross-validation. Recent work has shown that a large fraction of the performance of GBA methods can be attributed to the multifunctionality of genes [176]. It is not a priori clear exactly how the construction of the training and the test data sets affects the measured performance of a method. We show that Katz and CATAPULT outperform the state-of-the-art, as measured by standard cross-validation. Furthermore, we show that standard cross-validation is not always an appropriate yardstick for comparing the performance of methods, and that when an alternative method for cross-validation is used — measuring how well the methods do in case of singletons, simpler walk-based methods often achieve better performance than supervised learning counterparts. We also observe that the qualitative performance of the methods correlates better with the latter evaluation strategy. We evaluate the two proposed methods, and compare to state-of-the-art network-based gene-disease prediction approaches on two completely distinct sources of data, namely OMIM phenotypes and gene-drug interactions.

4.3 Results and Discussion

Conceptually, gene-disease association data can be thought of as a bipartite graph, where each gene and each disease is a node, and there is an edge between a gene node and a disease node if there is a known association between the gene and the disease. Similarly, we can form bipartite graphs from gene-phenotype association data of different species. By connecting a phenotype with a human gene if any *ortholog* of the human gene is associated with the phenotype, we obtain a bipartite network between human genes and phenotypes of different species. We can also obtain a phenotype-phenotype network for a given species, where a (weighted) edge (i, j) indicates that phenotype i is "similar" to phenotype j. Adding a gene-gene interaction network completes a *heterogeneous* network of human genes and phenotypes in a wide variety of species. It is straight-forward to define analogous heterogeneous network for gene-drug interactions, by replacing gene-disease associations data of humans with gene-drug associations. More limited heterogeneous networks have been considered previously in the context of gene-disease predictions, like the network of protein-protein interactions and human diseases considered in [173], and in the context of gene-drug predictions [177]. In this way, the model organism phenotypes provide a new kind of links between genes, and we can leverage the independent information hidden in the model organism data for discovering novel associations between genes and human diseases or drugs. A visualization of the heterogeneous network consisting of gene-gene network and gene-phenotype networks of a few model species is presented in Figure 4.1.

In this setting, it is natural to view the problem of predicting gene-phenotype

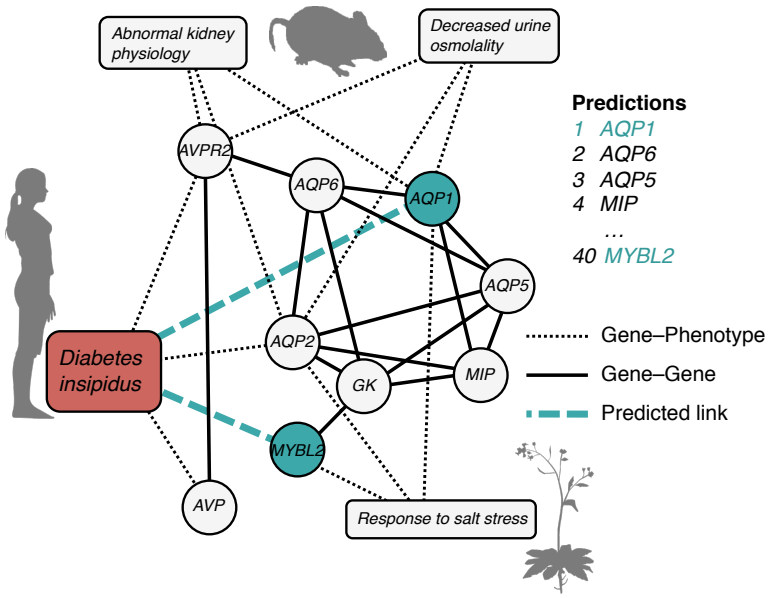


Figure 4.1: **The combined network in the neighborhood of a human disease.** The local network around the human disease diabetes insipidus and two genes highly ranked by CATAPULT, *AQP1* (top ranked candidate) and *MYBL2* (ranked as number 40). *AQP1* is ranked higher than *MYBL2* because there are more paths from diabetes insipidus to *AQP1* than to *MYBL2*, both through model organism phenotypes and through the gene–gene network. Only genes and phenotypes that are associated to both diabetes insipidus and the predicted genes *AQP1* and *MYBL2* are shown.

associations as a problem of finding similarities between nodes in a heterogeneous graph. Posing the problem in this way comes with the significant advantage that we can leverage a large body of work in machine learning and network analysis that deals with the problem of finding similar nodes in a graph (see, for example, [178, 179] and references therein). In particular, we adapt the Katz method from [178] to the heterogeneous setting. As an extension of this work, we also introduce a supervised learning framework, CATAPULT. CATAPULT learns the importance of features associated with node pairs, where the features are derived from walk-based similarity measures between nodes.

Katz on the heterogeneous network

Katz is a graph-based method for finding nodes similar to a given node in a network[175]. It has been shown to be successful for predicting friends in social network [178]. In this paper, we show the effectiveness of the method to recommend genes to given phenotype or drug. Suppose we are given an undirected, unweighted graph with a (symmetric) adjacency matrix A, where $A_{ij} = 1$ if node i and node j are connected, and $A_{ij} = 0$ otherwise. One way to find the similarity between two (not necessarily connected) nodes i and j is to count the number of *walks* of different lengths that connect i and j. This has a natural connection to matrix powers since $(A^l)_{ij}$ is exactly the number of walks of length l that connect i to j. $(A^l)_{ij}$ gives a measure of similarity between the nodes i and j. We want to obtain a single similarity measure that summarizes the similarities suggested by different walk lengths. For example, we could choose any sequence β_l of non-negative

coefficients and define the similarity

$$S_{ij} = \sum_{l=1}^{k} \beta_l (A^l)_{ij},$$

where β is a constant that dampens contributions from longer walks. In matrix notation, the similarity matrix S may be written as:

$$S = \sum_{l=1}^{k} \beta_l A^l \,. \tag{4.1}$$

As observed by the survey article [179], we can regard S as a matrix function $F(A)$ where F is defined through the series expansion in (4.1), and we may allow $k \to \infty$. Specific choices for β_l yield a variety of concrete similarity measures. A choice of $\beta_l = \beta^l$ for some β leads to the well-known *Katz* measure [175]:

$$S^{katz} = \sum_{l \geq 1} \beta^l A^l = (I - \beta A)^{-1} - I \,, \tag{4.2}$$

where β is chosen such that $\beta < \frac{1}{\|A\|_2}$. In the case where the connections in the graph are weighted such that A_{ij} is the strength of the connection between nodes i and j, we can generalize the idea of walks using this matrix framework, by simply using the weighted adjacency matrix instead of the binary matrix. Different ways of constructing the matrix A together with the appropriate normalizations of the matrix lead to methods of the type used by PRINCE [172], RWRH [173], and GeneMANIA [73], and by the famous PageRank algorithm used for web page ranking [180]. However, we do not necessarily have to consider sums over infinite path lengths. Paths of shorter lengths often convey more information about similarity between a given pair of nodes, and contributions from longer paths become insignificant. This suggests that we can consider a finite sum over path lengths,

and typically small values of k ($k = 3$ or $k = 4$) are known to yield competitive performance in the task of recommending similar nodes[181].

Let G denote gene-gene network, let P denote the bipartite network between genes and phenotypes, and let Q denote the phenotype-phenotype network. In particular, $P = [P_{Hs} \;\; P_S]$ is a composite of gene-disease network of humans, written P_{Hs}, and gene-phenotype networks of other species, written P_S. Similarly,

$$Q = \begin{bmatrix} Q_{Hs} & 0 \\ 0 & Q_S \end{bmatrix},$$

where Q_{Hs} is the similarity matrix of human diseases, and Q_S is that of phenotypes of other species. In our experiments, we set $Q_S = 0$, since we do not have information about similarity between phenotypes of other (non-human) species. The construction of the matrices G, P and Q_{Hs} will be discussed in detail in the Methods section. We form a *heterogeneous* network over the gene and phenotype nodes, similar to RWRH (which we will review briefly in the Methods section). The adjacency matrix of the heterogeneous network may be written as:

$$C = \begin{bmatrix} G & P \\ P^\top & Q \end{bmatrix}, \tag{4.3}$$

Recall the general formula for the Katz similarity measure when specialized to the combined matrix C:

$$S^{Katz}(C)_{ij} = \sum_{l=1}^{k} \beta^l (C^l)_{ij}, \tag{4.4}$$

Note that for smaller values of β, higher order paths contribute much less. It has been shown that restricting the sum to a small k, i.e. a few higher order paths works well in practice, in network link prediction and recommender systems[181].

Letting $k = 3$, the block of Katz score matrix $S^{Katz}(C)$ corresponding to similarities between gene nodes and human disease nodes, written $S^{Katz}_{Hs}(C)$, can be expressed as:

$$S^{Katz}_{Hs}(C) = \beta P_{Hs} + \beta^2(GP_{Hs} + P_{Hs}Q_{Hs}) + \beta^3(PP^TP_{Hs} + G^2P_{Hs} + GP_{Hs}Q_{Hs} + P_{Hs}Q^2_{Hs}) \tag{4.5}$$

where P_{Hs} and Q_{Hs} denote the gene-phenotype and phenotype-phenotype networks of humans respectively. We use Equation (4.5) to compute scores for Katz method in experiments.

In case of the drugs data set, we use the gene-drug network D, instead of P_{Hs} in Equation (4.5). We do not have similarity information for drugs, and so we set $Q = 0$ for experiments on drug data set. Nonetheless, we use phenotype information from multiple species (in the composite matrix $P = [D \;\; P_S]$) in order to infer similarities between gene and drug nodes.

CATAPULT: A supervised approach to predicting associations

The fixed choice of parameters involved in the Katz and random walk based approaches, as in Equation (4.4), provides a reasonable initial approach. However, to improve performance we would like to *learn* the weights based on the heterogeneous network itself. To this end, we frame the problem of predicting potential gene-phenotype associations as a supervised learning problem, in which we want to learn a classifier function whose input space consists of gene-phenotype pairs and output is a score for each gene-phenotype pair. In particular, by appropriately defining the feature space for gene-phenotype pairs, we will see that learning a

classifier in the constructed feature space is tantamount to learning coefficients for Katz on the heterogeneous network computed as in Equation (4.5). Our learning strategy is guided by the following two key characteristics of our data set:

1. For each phenotype, we only have a partial list of the associated genes. That is, we only know of positive associations; we do not have negative associations available to us.

2. There is a large number of unlabeled gene-phenotype pairs with the prior knowledge that most of them are, in fact, negative associations.

Classical supervised learning methods require both positive and negative examples, and therefore fall short in our case. Positive-Unlabeled learning (PU learning for short) methods are natural for this setting. The general idea of PU learning methods is to identify a set of negatives from the unlabeled examples and train a supervised classifier using the positives and the identified negatives. Liu et al[182] study different ways of choosing negatives from unlabeled examples. Biologists believe that only a few of the large number of *unobserved* associations are likely to be positive. A random sample is likely to consist mostly of negatives, which suggests that we could randomly choose a set of examples and use the random sample as "negative" examples to train a supervised classifier. As the examples are *not* known be negative, it may be helpful to allow the classifier to not heavily penalize the mistakes on "negatives" in the training phase. We therefore learn a *biased* support vector machine classifier using the positive associations and a *random* sample of unlabeled associations. Recently, Mordelet et al[183] also used a random

sample of unlabeled examples as negative sample to train a biased support vector machine against a set of known positives. The support vector machine is biased in the sense that false negatives (known positives classified as negatives) are penalized more heavily than the false positives ("negatives" classified as positive). The bias makes sense because the positive examples are known to be positive, while the negatives were arbitrary and hence false positives are not to be penalized too heavily. Note that, in principle, we could use any PU learning method (for instance, the weighted logistic regression model proposed in [184]) to obtain a classifier for gene-phenotype pairs.

Recent work by [183] uses the *bagging* technique to obtain an aggregate classifier using positive and unlabeled examples. In this approach one draws a random bootstrap sample of a few unlabeled examples from the set of all unlabeled examples and trains a classifier treating the bootstrap sample as negatives along with the positive examples. Bagging helps to reduce the variance in the classifier that is induced due to the randomness in the "negative" samples. Let T be the number of bootstraps, let $\mathcal{A}$ be the set of positives (*i.e.* gene-phenotype pairs that correspond to known associations), let n_+ denote the number of examples in $\mathcal{A}$, and let $\mathcal{U}$ denote the set of unlabeled gene-phenotype pairs. We train a *biased* SVM, where we use a penalty C_- for false positives and relatively larger penalty C_+ for false negatives. Let $\Phi : \mathcal{G} \times \mathcal{P} \rightarrow \mathbb{R}^d$ denote a feature map for the gene-phenotype pairs, where $\mathcal{G}$ is the set of genes and $\mathcal{P}$ is the set of phenotypes (of multiple species). We will discuss toward the end of the section how walk-based measures like Katz can be used to obtain an embedding.

CATAPULT (Combining dATa Across species using Positive-Unlabeled Learning Techniques) uses the biased SVM framework to classify gene-phenotype pairs of humans and multiple other species with a single training phase, thereby making the best use of the relation between different phenotypes. The bagging algorithm that trains and combines several biased SVM classifiers used by CATAPULT is as follows:

initialize $\theta = 0$ and $n(x) = 1, \forall x \in \mathcal{U}$.

for $t = 1, 2, \ldots, T$:

1. Draw a bootstrap sample $\mathcal{U}_t \subseteq \mathcal{U}$ of size n_+.

2. Train a linear classifier θ_t using the positive training examples $\mathcal{A}$ and $\mathcal{U}_t$ as negative examples by solving:

$$\begin{aligned} \min_{\theta' \in \mathbb{R}^d} \quad & \|\theta'\|^2 + C_- \textstyle\sum_{i \in \mathcal{U}_t} \xi_i + C_+ \sum_{i \in \mathcal{P}} \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \forall i \in \mathcal{A} \cup \mathcal{U}_t, \\ & \langle \Phi(x_i), \theta' \rangle \geq 1 - \xi_i, \forall i \in \mathcal{A}, \text{ and} \\ & -\langle \Phi(x_i), \theta' \rangle \geq 1 - \xi_i, \forall i \in \mathcal{U}_t. \end{aligned} \tag{4.6}$$

3. Update: $\theta \leftarrow \theta + \theta_t$.

4. For any $x \in \mathcal{U} \setminus \mathcal{U}_t$ update: $n(x) \leftarrow n(x) + 1$.

return $s(x) = \langle \theta, \Phi(x) \rangle / n(x), \forall x \in \mathcal{U}$.

We train a biased SVM given in equation (4.6) during each iteration using all the known positive examples in $\mathcal{A}$ and a randomly chosen set of "negatives" $\mathcal{U}_t \subseteq \mathcal{U}$. Positive and negative examples may not be linearly separable, and the standard way is to penalize based on how far an example is from meeting its margin requirement, through the use of *slack* variables ξ_i. The scoring function for iteration t is proportional to the distance of the point x from the hyperplane and is given by the standard dot product,

$$\langle \theta_t, \Phi(x) \rangle$$

where θ_t is the normal to the hyperplane learned using the random bootstrap at the tth iteration and $\Phi(x)$ is the feature vector corresponding to x. For small number of boostraps, say T in the range 10-100, $n(x) = T$ for most of the unlabeled examples and thus the procedure in effect scores (most of the) unlabeled examples using the average hyperplane $\frac{1}{T}\sum_t \theta_t$. We set $T = 30$ in our experiments. Recall that, in our framework, an instance x corresponds to a gene-phenotype *pair*. In contrast to the traditional SVM classifiers that classify a pair as positive or negative based on the sign of $\langle \theta_t, \Phi(x) \rangle$, we use the value as a score under the assumption that the further a point is on the positive side of the hyperplane, the more likely it is to be a true positive.

Parameters. In Equation (4.6), $C_+ \gg C_-$ are the penalties on misclassified positives and negatives respectively. The weights control the relative widths of the margins on either sides of the hyperplane. As C_+ increases from 0 to ∞, the margin on the side of the positive examples shrinks, and as $C_+ \to \infty$, the classifier makes

no mistake on the positive examples. The ratio C_+/C_- determines the "weight" of a positive example, and we want this to be a very high value. In our experiments, we set $C_- = 1$ and $C_+ = 10$, which is found to be the best by cross-validation. The cross-validation is done as follows: We split the positive examples into 5 folds. Using each fold as test set in turn, we do the bagging procedure with the remaining 4 folds as training positives $\mathcal{A}$. For a given setting of C_+ and C_-, we obtain the average recall of the final bagged classifier on the hidden test set, i.e. fraction of number of true positives identified in the top k predictions, where k is the number of positive examples in the test set. We choose the values of C_+ and C_- that achieve the highest average recall in cross-validation[1].

Features derived from hybrid walks. Before applying any supervised machine learning approach, we need to construct *features* for gene-phenotype pairs. The features that we use are all based on paths in the combined heterogeneous network. Recall that in the Katz measure, the weights for combining the contributions from walks of different lengths is fixed beforehand. We observe from Equation (4.5) that, for a given length of walk, there are multiple ways of obtaining hybrid walks, as given by the terms in the series. For a given gene-phenotype pair, different walks of the same length, and walks of different lengths can be used as features for the pair. Thus learning a biased SVM provides an efficient way to learn the weights, and could help improve on the prediction performance over a particular choice of weights, say, $(\beta, \beta^2, \beta^3, \dots)$ as in Katz. Clearly, the dimensionality of the feature

[1]Fixing $C_- = 1$, $\log_{10} C_+$ was varied in the range $-3, -2, -1, 0, 1, 2, 3$ and $\log_{10} C_+ = 1$ was found to be the best.

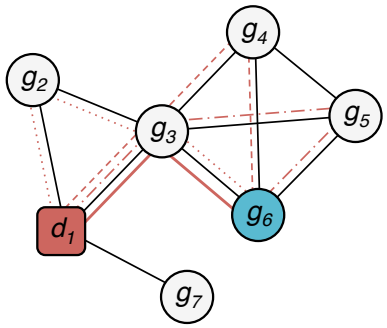


Figure 4.2: **Katz features are derived by constructing walks of different kinds on the graph.** In the figure above, the disease node d_1 is connected to the gene node g_6 by one walk of length 2 (solid red line) and three walks of length 3 (dotted, dashed and dashdotted red lines). This can be quickly calculated from the adjacency matrix C of the graph: If $C_{ij} = 1$ when there is a link between nodes i and j, and 0 otherwise, the number of paths of length n between genes i and j is $(C^n)_{ij}$. In the example above, $\left(C^2\right)_{16} = 1$ and $\left(C^3\right)_{16} = 3$.

space is exponential in k, length of the walk, and makes us vulnerable to the curse of dimensionality because the examples are limited. However, taking a cue from the fact that the weights of increasing walk lengths need to be heavily damped, we ignore higher order terms and thereby keep the dimensionality of the feature space small. We further decompose the features to distinguish different species, which enables learning the contribution of each species to the prediction. The complete set of features used by CATAPULT and the corresponding weights learned are listed in Table 4.5. We also observe from our experiments that using species-wise features not only lends interpretability but also improves the accuracy of the predictions, as compared to combining features corresponding to same walk lengths. Figure 4.2 demonstrates simple walk-based features derived from the heterogeneous network.

Functional data outperforms protein-protein interactions

To see how the Katz and CATAPULT methods compare to the state-of-the-art, we measured their recovery of genes using a cross-validation strategy similar to the one in [174], on two different data sets, gene-disease associations from the Online Mendelian Inheritance in Man (OMIM, [185]), and a recent drug-gene interaction data set from [177]. These data sets can both be thought of as a large collection of gene-trait pairs, either as gene-disease pairs for the OMIM data, or target-drug pairs for the drug data set.

We compared Katz and CATAPULT to four recent methods:

1. The recently proposed **ProDige** method from [174], which is a support vector machine based method that calculates similarity scores for gene pairs using a wide variety of information sources including 21 different gene-gene functional interaction networks and phenotype similarities.

2. **RWRH** from [173], which, like Katz uses walks on a heterogeneous gene-disease graph to prioritize genes. It differs from the Katz method chiefly in how the heterogeneous network is normalized. We discuss the relationship in more detail in the Supplementary Material.

3. We include **PRINCE** [172] for completeness, since it is the state-of-the-art to which both RWRH and ProDiGe were compared.

4. Finally, some recent work [176] has shown that simply by ranking based on the degree centrality of a gene (how often it interacts with other genes, or is

involved in diseases) can be a very competitive ranking strategy. We therefore predict genes for diseases (or drugs) using a simple degree-based list, where all genes are ranked by how many diseases (drugs) they are known to be connected to, *regardless* of which disease (drug) the predictions are made for.

For cross-validation, we use the same testing framework as the one used in [174]: We split the known gene-trait pairs into three equally sized groups. We hide the associations in one group and run our methods on the remaining associations, repeating thrice to ensure that each group is hidden exactly once. For each trait in our data set, we order all the genes by how strongly the method predicts them to be associated with the trait. Finally, for every gene-trait pair (g, t) in the hidden group we record the rank of the gene g in the list associated with trait t.

The results are presented in Figure 4.3. Under this evaluation method, both Katz and CATAPULT, which make use of much more extensive data sets than the other methods, are quite likely to recover the hidden gene among the top 100 genes. As can be seen from Figure 4.3, Katz and CATAPULT perform better than *any* of the previously studied state-of-the-art gene-disease association methods for the OMIM data set. CATAPULT also performs well on the drug data set, ranking the hidden gene 14th or lower a remarkable 50% of the time. RWRH, which like Katz and CATAPULT is a walk based method that allows paths through the gene-disease (or, for the drug data set, gene-drug) network, also does quite well.

ProDiGe allows sharing of information between phenotypes using the similarities between OMIM phenotypes, and also integrates a wide variety of functional

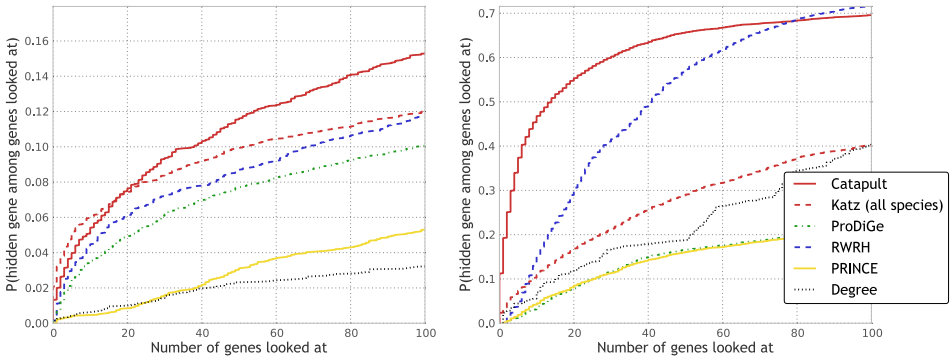


Figure 4.3: **Empirical cumulative distribution function for the rank of the withheld gene under cross-validation.** Left panel corresponds to evaluation of OMIM phenotypes, and the right corresponds to drug data. Katz and CATAPULT methods use all species information, and **HumanNet** gene network. PRINCE and RWRH methods are implemented as proposed in [172] and [173] respectively, using the **HPRD** gene network. ProDiGe method is implemented as discussed in Methods section. CATAPULT (solid red) does much better across the data sets under this evaluation scheme. In general, the methods get high precision rates in case of the drug data. PRINCE method that does not allow walks through species phenotypes, and OMIM phenotypes in particular, performs much worse than other random-walk based methods. ProDiGe allows sharing of information between phenotypes using the similarities between OMIM phenotypes and performs reasonably well, whereas there is no such sharing possible in case of the drug data due to the absence of drug similarities. Simple degree-based method performs poorly in general. ProDiGe and PRINCE essentially use only the gene network information in case of the drug data.

information in a supervised machine learning framework and performs reasonably well on the OMIM data set. The PRINCE method, which allows some sharing of information between OMIM diseases that are phenotypically similar, performs worse than the other random-walk based methods. Since we have no similarity information available for the drug data, ProDiGe and PRINCE essentially use only the gene similarity information in the drug data case. Notice that the simple degree-based method does the worst of all methods in case of OMIM dataset, which suggests that recommendations given by walk-based methods are more relevant and differ significantly from simple ranking by number of known associations.

To see if the improvement in performance of Katz and CATAPULT stems from the more extensive network used, or, in CATAPULT's case, the increased sophistication of the machine learning method, we evaluated network based RWRH and PRINCE methods using the more extensive HumanNet network instead the HPRD network originally used. As can be seen in Figure 4.4 CATAPULT still does better than the previous state-of-the-art using this cross-validation framework, consistently in both the OMIM and drug data sets.

Top candidates are enriched for highly connected genes

To get a qualitative view of how the connectedness of genes influences the rankings, we plotted the degree distribution of the OMIM and drug datasets in Figure 4.5, and compared the results with the list of top candidates from CATAPULT (see Table 4.1) and Katz (see Table 4.2).

The results for CATAPULT all seem very reasonable, from a biological stand-

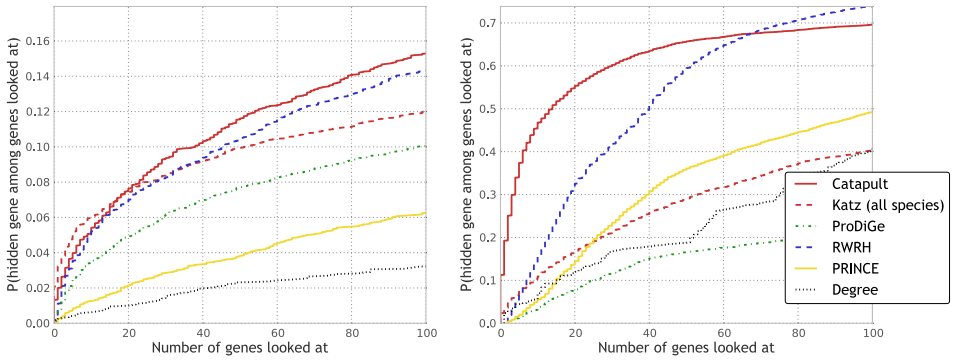


Figure 4.4: **Comparison of performances using only HumanNet.** Empirical cumulative distribution function for the rank of the withheld gene under cross-validation. Left panel corresponds to evaluation of OMIM phenotypes, and the right corresponds to drug data. Katz and CATAPULT methods use all species information, and all the methods use **HumanNet** gene network. PRINCE and RWRH methods are implemented as proposed in [172] and [173] respectively, but using **HumanNet** gene network. ProDiGe method is implemented as discussed in Methods section. Again, as in Figure 4.3, CATAPULT (solid red) does the best. An important observation to be made from the Figure is that PRINCE and RWRH methods perform relatively much better than in Figure 4.3, i.e. when HPRD network is used. (Note that there is no change to ProDiGe, Katz and CATAPULT methods; they have identical settings as in Figure 4.3).

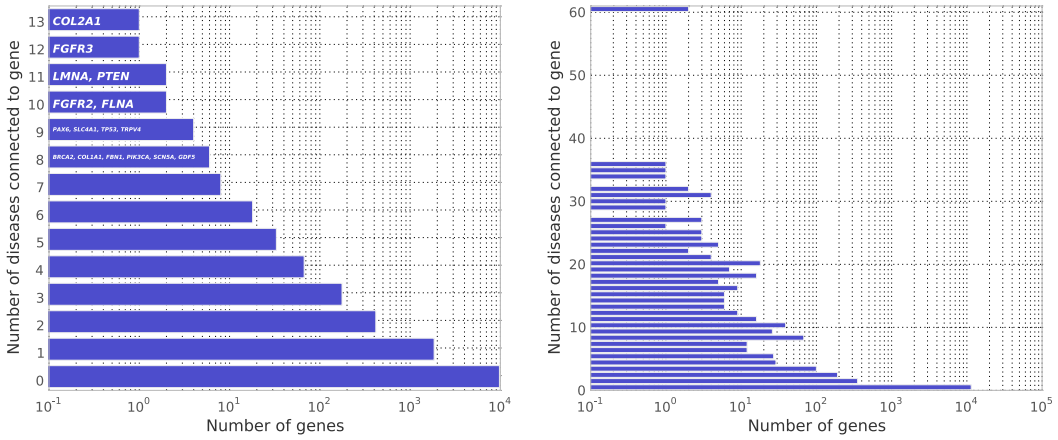


Figure 4.5: **Distribution of the number of known phenotype associations per gene,** in OMIM diseases (left) and drugs (right).

Table 4.1: **Top 10 predictions by** CATAPULT **for the eight OMIM phenotypes with most known causal genes.** Any gene which is among the top 10 candidates for more than one disease is marked in bold. CATAPULT does make a great number of very reasonable predictions as observed below. For example, it seems quite likely that both insulin receptor (INSR, 3643) and insulin (INS, 3630) should be associated with insulin resistance, and that many growth factor receptors have been associated with various cancers.

Leukemia **MIM:601626**	**Alzheimer disease** **MIM:104300**	**Insulin resistance** **MIM:125853**	**Prostate cancer** **MIM:176807**
FGFR3 (2261)	*ACE2* (59272)	***INSR*** (3643)	***TP53*** (7157)
FGFR2 (2263)	*COL1A1* (1277)	*INS* (3630)	***RB1*** (5925)
KRAS (3845)	*COL1A2* (1278)	***PTEN*** (5728)	***CTNNB1*** (1499)
TP53 (7157)	***KRAS*** (3845)	***TP53*** (7157)	***BRCA1*** (672)
EGFR (1956)	***EGFR*** (1956)	***CTNNB1*** (1499)	***KRAS*** (3845)
FGFR1 (2260)	***TP53*** (7157)	***KRAS*** (3845)	***PIK3CA*** (5290)
PTPN11 (5781)	*AGT* (183)	***AKT1*** (207)	***AKT1*** (207)
CTNNB1 (1499)	*PLAT* (5327)	***CREBBP*** (1387)	***INSR*** (3643)
INSR (3643)	*APOE* (348)	***EGFR*** (1956)	***NRAS*** (4893)
CREBBP (1387)	*PTGS2* (5743)	***PIK3CA*** (5290)	*RAD51* (5888)

Schizophrenia **MIM:181500**	**Breast cancer** **MIM:114480**	**Gastric cancer** **MIM:137215**	**Colorectal cancer** **MIM:114500**
BDNF (627)	***PTEN*** (5728)	***FGFR3*** (2261)	***KRAS*** (3845)
NRG1 (3084)	***RB1*** (5925)	***FGFR1*** (2260)	***PTEN*** (5728)
CBS (875)	***NRAS*** (4893)	***NRAS*** (4893)	***CTNNB1*** (1499)
NOS2 (4843)	***BRCA1*** (672)	***HRAS*** (3265)	***HRAS*** (3265)
MTR (4548)	***HRAS*** (3265)	***EGFR*** (1956)	***CREBBP*** (1387)
HTR2C (3358)	***INSR*** (3643)	*ERBB3* (2065)	***RB1*** (5925)
HTR2B (3357)	***CTNNB1*** (1499)	***CTNNB1*** (1499)	***FGFR3*** (2261)
SLC6A4 (6532)	***EGFR*** (1956)	*BRAF* (673)	***INSR*** (3643)
FGFR2 (2263)	***FGFR3*** (2261)	***PTEN*** (5728)	***EGFR*** (1956)
MAT1A (4143)	***FGFR2*** (2263)	***TP53*** (7157)	***FGFR2*** (2263)

Table 4.2: **Top 10 predictions by Katz for the same phenotypes as in Table 4.1.** Any gene which is among the top 10 candidates for more than one disease is marked in bold. The Katz method shows a weaker link between the number of diseases previously associated with a gene and its presence in the list, while still giving a number of very likely candidates.

Leukemia **MIM:601626**	**Alzheimer disease** **MIM:104300**	**Insulin resistance** **MIM:125853**	**Prostate cancer** **MIM:176807**
IL3 (3562)	*APLP2* (334)	*INS* (3630)	***BRCA1*** (672)
SOCS1 (8651)	*HSPA8* (3312)	***AKT1*** (207)	*TP53* (7157)
GRB2 (2885)	*CTSB* (1508)	***INSR*** (3643)	*RAD51* (5888)
NOP2 (4839)	*LRP1* (4035)	***GRB2*** (2885)	***EGFR*** (1956)
CSF2RB (1439)	*NID1* (4811)	***IGF1R*** (3480)	*ATM* (472)
PPM1L (151742)	*APOE* (348)	***CTNNB1*** (1499)	***AKT1*** (207)
PTPN6 (5777)	*BDKRB2* (624)	*CREBBP* (1387)	*MAX* (4149)
MYH11 (4629)	*PLAUR* (5329)	***PIK3CA*** (5290)	***CDK1*** (983)
PPM1E (22843)	*APLP1* (333)	*TYK2* (7297)	***PIK3CA*** (5290)
PPM1B (5495)	*CAV1* (857)	*GPD1* (2819)	*CSNK2A1* (1457)

Schizophrenia **MIM:181500**	**Breast cancer** **MIM:114480**	**Gastric cancer** **MIM:137215**	**Colorectal cancer** **MIM:114500**
DRD2 (1813)	***BRCA1*** (672)	***GRB2*** (2885)	***PTEN*** (5728)
AHCY (191)	***IRS1*** (3667)	***EGFR*** (1956)	***CTNNB1*** (1499)
ADRA2B (151)	*MRE11A* (4361)	*NRAS* (4893)	***CDK1*** (983)
XRN2 (22803)	***INSR*** (3643)	***IRS1*** (3667)	*GSK3B* (2932)
MAT1A (4143)	*CHEK1* (1111)	***MAPK1*** (5594)	*CDC20* (991)
MAT2A (4144)	*ATR* (545)	*PTPN11* (5781)	*5111* (5111)
CHI3L2 (1117)	***PTEN*** (5728)	*HRAS* (3265)	*EGF* (1950)
TSNAX (7257)	***MAPK1*** (5594)	*MAP2K2* (5605)	*PTTG1* (9232)
DDC (1644)	*MAPK3* (5595)	*MAP2K1* (5604)	***IGF1R*** (3480)
MAOB (4129)	*UBE2I* (7329)	*SOS1* (6654)	*FOXO3* (2309)

point. For example, CATAPULT identifies *APOE*, which even though is not linked to "Susceptibility to Alzheimer's disease" OMIM record (MIM:104300), is well known to be associated with Alzheimer's disease and is associated with two other OMIM records involving Alzheimer's (MIM:104310 and MIM:606889). BRCA1 is associated with "Breast-ovarian cancer, familial 1" (MIM:604370), not the record we show in Table 4.1 ("Breast cancer, susceptibility to", MIM:114480), even so, it is ranked very highly among the candidate genes for breast cancer. Many of the other candidate genes listed are similarly very likely to be involved in the etiology of the diseases, like *TP53* and *KRAS* for many cancers. Indeed, what might be the most surprising about the results is how completely non-surprising they seem. Furthermore, there is a very high degree of overlap between the top predictions. Indeed, almost all the top 10 candidate genes for the eight diseases shown are shared between at least two of the eight diseases for CATAPULT. Moreover, when studying the results for the same diseases for ProDiGe, given in [174], we see the same pattern as we see for CATAPULT – a strong enrichment for genes that are already known to be associated with many diseases.

In contrast, the results for Katz (Table 4.2) contain comparatively fewer of these very obvious predictions, and a much lower degree of overlap between the top predictions. There is still a certain number of predictions shared, particularly between the different cancers and type 2 diabetes. However, the predictions made seem to reflect the relevance of a gene to the specific disease more than the overall likelihood that a gene is associated with *any* disease.

Validation on singletons highlights methods that detect novelty

The cross-validation evaluation shown in Figures 4.3 and 4.4 clearly shows that CATAPULT is better at recapitulating the genes known to be involved in a disease than any of the other methods. However, recapitulation of previously known results is rarely the goal in biology. We therefore seek a measure that would reflect how suited a method is for correctly identifying novel associations.

There are two ways in which one could envision for doing this in a cross-validation framework – either one could hide *all* associations between a given gene and diseases, thereby hoping to put it on equal footing with genes still unstudied, or one restricts the cross-validation to genes that are only associated with a single phenotype. There are clear advantages to both approaches. The former approach allows us to do validation on a larger set, namely all known gene-disease associations, and thereby reach stronger statistical strength. The latter approach has more subtle, but in our opinion greater, advantage. The biases that favor already well studied genes are not only present in the gene-disease association data, but also in the data that gives rise small differences between genes that have been studied well and poorly characterized genes. By only looking at the least studied genes in our data set for which we do have known gene-disease associations, we can minimize the risk that any signal that we detect is merely some general characteristic of well studied genes, and instead actually measure how well a method can detect truly novel gene disease associations.

We tested all the methods using cross-validation restricted to genes with only a single disease (or drug) association. The results are presented in Figure 4.6.

CATAPULT does worse on singleton phenotypes and drugs, as compared to Figure 4.4 (that uses the same setting for all the methods). PRINCE and ProDiGe methods are consistent with (and sometimes perform slightly better than) the three-fold cross-validation evaluation. RWRH and Katz perform better than the supervised learning methods ProDiGe and CATAPULT in this evaluation sheme. The fact that PRINCE performs the best on singletons in case of drug data is surprising, given that the only information it uses is the HumanNet gene network. Simpler random-walk based methods in general perform better than the supervised counterparts, and do so consistently in two completely distinct data sets. Furthermore, we find that the qualitative results of the methods (Tables 4.1 and 4.2) are in line with the observed performance of the methods on singleton genes.

Conclusions

We have proposed two methods for inferring gene-phenotype associations, Katz and CATAPULT. Katz is motivated by social network link prediction and CATAPULT is a supervised extension to Katz which learns the weights for walks that have different lengths and that involve different kinds of data. While CATAPULT significantly outperforms other state-of-the-art gene-phenotype association methods using a conventional cross-validation evaluation strategy, such a cross-validation strategy does not reflect the properties of a scenario in which one wants to predict *novel* gene-phenotype associations involving less studied genes.

To address such cases, we propose a cross-validation approach restricted to relatively little studied genes. In this framework the Katz method and the related

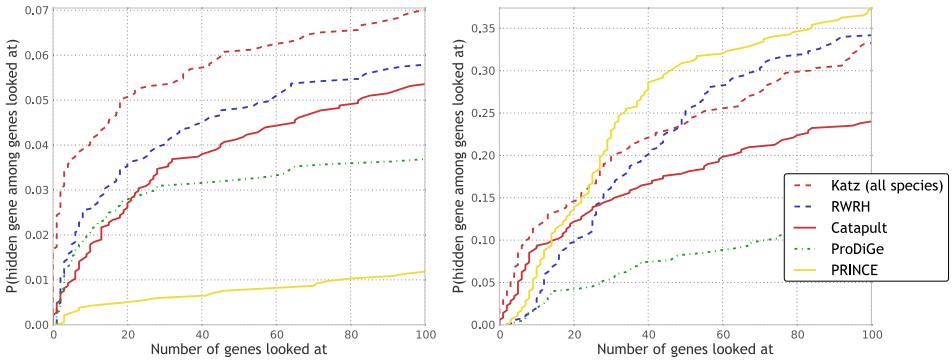


Figure 4.6: **Empirical cumulative distribution function for the rank of the withheld gene, under evaluation of singleton phenotypes and drugs.** Left panel corresponds to evaluation of OMIM phenotypes, and the right corresponds to drug data. Katz and CATAPULT methods use all species information, and all the methods use **HumanNet** gene network. PRINCE and RWRH methods are implemented as proposed in [172] and [173] respectively, but using **HumanNet** gene network. ProDiGe method is implemented as discussed in Methods section. CATAPULT (solid red) does worse on singleton phenotypes and drugs, as compared to Figure 4.4 (that uses the same setting for all the methods). PRINCE and ProDiGe methods are consistent with (and sometimes perform slightly better than) the full cross-validation evaluation. RWRH and Katz perform better than the supervised learning methods ProDiGe and CATAPULT in this evaluation sheme. The fact that PRINCE performs the best on singletons in case of drug data is surprising, given that the only information it uses in the HumanNet gene network.

RWRH and Prince methods do better than CATAPULT, indicating that if the objective is to find new gene-disease or gene-drug associations involving genes not yet well studied, these approaches are more appropriate.

We therefore conclude that comparisons of gene-phenotype methods do not necessarily lead to a simple ordering from best to worst. A method like CATAPULT is clearly preferable when the goal is to recapitulate the currently known gene-phenotype associations. On the other hand, in cases where a researcher wants to find new directions for research or find previously unknown biology, a method like Katz, which does better when tested on genes only associated with a single disease, is clearly preferable to a method like CATAPULT, which emphasises genes that are "important" in general. In the future, it is therefore important that descriptions of new gene-phenotype association methods include a careful discussion on how the method is intended to be used.

4.4 Materials and Methods

Gene Networks

We use two sources of gene-gene interactions in our experiments.

1. **HumanNet**: A large-scale functional gene network which incorporates multiple data sets, including mRNA expression, protein-protein interactions, protein complex data, and comparative genomics (but not disease or phenotype data) obtained from [21]. HumanNet contains 21 different data sources, which are combined into one integrated network using a regularized regression scheme trained on KEGG pathways.

2. **HPRD** network [186]: Most of the published work on predicting gene-disease associations [170–173, 187] use the HPRD network. The network data was downloaded from [169]. The edges in the HPRD network are unweighted, and the network is much sparser than HumanNet. In particular, the HPRD network has 56,661 associations compared to 733,836 (weighted) associations for HumanNet.

Phenotypes from other (non-human) species

We collected gene-phenotype associations from literature and public databases for eight different (non-human) species: plant (*Arabidopsis thaliana*, from TAIR [164]), worm (*Caenorhabditis elegans* from WormBase [188] and [159]), fruit fly (*Drosophila melanogaster* from FlyBase [189]), mouse (*Mus musculus* from MGD [190]), yeast (*Saccharomyces cerevisiae* [9, 191–193]), *Escherichia coli* [161], zebrafish (*Danio rerio* from ZFIN [194]), and chicken (*Gallus gallus* from GEISHA [195]). We determined orthology relationships between genes in model species and human using INPARANOID [112]. Detailed description on the extraction of most datasets can be found in [11] and the resulting dataset has been summarized in Table 4.3.

E. coli phenotypes were obtained from the file 'coli_FinalData2.txt' on May 20, 2011 [161]; we sorted each gene's phenomic profile by score, taking both the top and bottom forty conditions and assigning them to the gene. Thus, we considered each condition to be a phenotype, and the genes associated with that phenotype were those genes whose growth was most affected (either positively or negatively) in the corresponding condition. As a proxy for chicken phenotypes, tissue specific

mRNA expression patterns were derived from GEISHA *in situ* hybridization annotations, which were kindly provided in XML format on June 24, 2011. Genes were sorted into multiple bins by stage, by location, and by location and stage together. If there were more than fifty genes in a specific location and more than three at a specific stage at that location, a new phenotype was created ("*anatomical location* at stage *x*"); regardless, each location became a phenotype. Worm phenotypes from [159] were divided into two datasets, 'green-broad' and 'green-specific', based on the broad and specific phenotypes presented in that work. GO biological processes from TAIR and ZFIN were processed in the same manner. We kept only those annotations with evidence codes IMP, IDA, IPI, IGI, TAS, NAS, IC, and IEP. For TAIR, we used 'ATH_GO_GOSLIM.txt', downloaded on August 23rd, 2010; and for ZFIN, we obtained GO biological processes from geneontology.org ('gene_association.zfin.gz') on April 26th, 2011.

Evaluation data

We perform experiments on two types of data sources:

- **OMIM Phenotypes**: We obtained new OMIM data from the Online Mendelian Inheritance in Man (OMIM) project [185] on August 11, 2011. OMIM phenotypes have become the standard data set for the evaluation of prediction of gene-disease associations[170–174, 187]. All the compared methods use similarities between phenotypes. We obtained similarities between OMIM phenotypes from [196].

- **Drug data**: This includes four benchmark data sets of Drug-Target interac-

tions in humans involving enzymes, ion channels, G-protein-coupled receptors (GPCRs) and nuclear receptors, first studied in [197]. Refer to Table 4.4 for statistics on the data sets. The data sets were made available by [177] and downloaded from [198].

Problem setup and Notation

Let $\mathcal{G}$ denote the set of human genes and for each species $i \in \mathcal{S} = \{Hs, At, Ce, Dm, Dr, Ec, Gg, Mm, Sc\}$, let $\mathcal{P}_i$ denote the set of phenotypes for the species. Refer Table 4.3 for descriptions of the species and a summary of the data sets. Also, let $\mathcal{D}$ denote the set of drugs (i.e. the four benchmark data sets mentioned in Table 4.4). For each species $i \in \mathcal{S}$, we constructed a gene-phenotype association matrix $P_i \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{P}_i|}$, such that $(P_i)_{gp} = 1$ if gene g is associated with phenotype p or 0 otherwise. For methods using multiple species, we used $P_S = [P_{At} \;\; P_{Ce} \;\; P_{Dm} \;\; \ldots \;\; P_{Sc}]$ and recall that $P = [P_{Hs} \;\; P_S]$ in Equation (4.5). Similarly, we constructed a drug-gene interaction matrix D using drugs data where $D_{gd} = 1$ if gene g is associated with drug d (note that d can be one of enzymes, ion channels, GPCRs or nuclear receptors) and $D_{gd} = 0$ otherwise. Using the two types of gene-gene interaction data HPRD and HumanNet, we constructed matrices $G^{HPRD} \in \{0,1\}^{|\mathcal{G}|\times|\mathcal{G}|}$, and $G^{HumanNet} \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{G}|}$ respectively. We constructed a phenotype-phenotype network $Q_{Hs} \in \mathbb{R}^{|\mathcal{P}_{Hs}|\times|\mathcal{P}_{Hs}|}$ (i.e. corresponding to humans) using OMIM phenotype similarities obtained from [196]. For experiments with drug data, we did not have access to any such similarity score for drug pairs, so we set the drug-drug network to 0. The same is the case for other species data as well, and we set the correspond-

ing entries in Q to be 0, both for the experiments with OMIM and for the drug data. Following the approach in [172], we apply a logistic transformation to the similarities Q_{Hs}, i.e. $L(x) = \frac{1}{1+\exp(cx+d)}$ where x represents an entry of Q_{Hs}. For setting c and d, see [172].

RWRH

Random Walks with Restart on Heterogeneous network (RWRH) is an algorithm for predicting gene-disease associations proposed by Li and Patra[173]. RWRH performs a random walk on a heterogeneous network of gene interactions (HPRD) and human diseases (we used OMIM phenotypes and the drug data described above). The method constructs a heterogeneous network using the G^{HPRD}, P_{Hs} and Q_{Hs} networks and runs a personalized PageRank computation, a popular choice for ranking documents and web pages, on the network. The random walk is started from a set of seed nodes, which for a phenotype p is the set of genes known to be associated with p, and gene nodes are ranked by the probability that a random walker is at a given gene, under the steady state distribution for the random walk. In particular, RWRH in [173] considers the following heterogeneous network:

$$C = \begin{bmatrix} \tilde{G} & \lambda P_{Hs} \\ \lambda P_{Hs}^{\top} & \tilde{Q}_{Hs} \end{bmatrix} \tag{4.7}$$

where $\tilde{G}$ is the gene-gene interactions matrix, with rows normalized by row-degree and scaled so that $\sum_j C_{ij} = 1$, and λ is the probability that the random walker jumps from a gene node to a phenotype node (or vice versa). In [173], P_{Hs} is the gene-disease association matrix corresponding to OMIM phenotypes, Q_{Hs} is the corresponding similarity matrix, and $\tilde{G}$ is derived from G^{HPRD}. Genes are ranked

for a given disease p using the steady state vector $\mathbf{s}$ given by :

$$\mathbf{s} = (1-\gamma)C^T\mathbf{s} + \gamma p_0 \tag{4.8}$$

where p_0 is the restart vector (indicator vector of the set of seed nodes known to be associated with p). In our experiments, we use OMIM phenotypes matrix P_{Hs} as well as the gene-drug interaction matrix D, and two types of gene-gene matrices to derive $\tilde{G}$. Recall that in the latter case, we do not have similarity information for drugs, and therefore we set drug-drug similarity matrix to 0. It is also straightforward to incorporate phenotype data from multiple species in the method, by replacing P_{Hs} with $P = [P_{Hs} \;\; P_S]$, analagous to our Katz method.

PRINCE

The PRINCE method, proposed by Vanunu *et al.* [172], is another graph-based method that can be thought of as a special case of RWRH. Here, the random walk is only over the gene interaction network instead of the heterogeneous network. The phenotype similarities are incorporated in the restart vector. The vector of scores computed by PRINCE for a given phenotype p can be expressed as

$$\mathbf{s}_{PRINCE} = (I - \gamma G)^{-1}\tilde{p} \tag{4.9}$$

where $\tilde{p}$ is the smoothed phenotype, i.e. $\tilde{p}_i = (Q_{Hs})_{qp}$ where q is the phenotype most similar to p. Note that, similarly, the scores computed by RWRH can be written succinctly as

$$s_{RWRH} = (I - \gamma C)^{-1}p \tag{4.10}$$

where C is defined in Equation (4.7). The absence of similarity information for other (non-human) species phenotypes and drugs renders direct extension of PRINCE to multiple species data inconsequential. We must emphasize here that PRINCE does not allow walks through the gene-phenotype interaction network or the phenotype-phenotype interaction network. As a result, availability of other species data becomes irrelevant when predicting genes for a given disease (or other drug data in case of predicting for a given drug).

ProDiGe

The ProDiGe method, proposed by Mordelet and Vert [174], makes use of positive-unlabeled learning and a multiple kernel learning framework to integrate information from multiple types of gene interaction data and phenotype similarities. Kernels are defined over pairs of genes and pairs of phenotypes, and the kernel value for a pair of gene-phenotype pairs is derived using the individual gene and phenotype kernels. In particular, let $K_{gene}(g, g')$ denote the kernel for genes, and $K_{phenotype}(p, p')$ denote that for phenotypes. Then, the kernel for the pairs $((g, p), (g', p'))$ is simply,

$$K_{pair}((g, p), (g', p')) = K_{gene}(g, g') \times K_{phenotype}(p, p') \tag{4.11}$$

The gene-phenotype pairs are then classified using a support vector machine using the constructed kernel. Note that the method proposed in [174] does not use any other species phenotype information, but only the OMIM phenotypes. In our experiments on OMIM phenotypes, we used the K_{gene} and $K_{phenotype}$ provided by the authors of [174]. For the drug data, we used a simple Dirac kernel for $K_{drug}(d, d')$,

i.e.

$$K_{drug}(d, d') = \begin{cases} 1 & \text{if } d = d', \\ 0 & \text{otherwise.} \end{cases}$$

For K_{gene} and $K_{phenotype}$, we used the kernels provided by Mordelet[2].

Implementation

We implemented all the methods in Matlab. Our implementation of CATAPULT can be downloaded from the web[3]. The training time for our method is essentially the time taken for constructing the features. Obtaining features for all gene-phenotype pairs takes about 20 minutes. Training and bagging biased SVMs are much faster, and take a few seconds per iteration on our cluster machines (2.8 GHz processor, 32GB RAM). We download the MATLAB code for Li and Patra's RWRH method[4], and the code for Mordelet and Vert's ProDiGe[5]. For PRINCE, we use MATLAB code kindly provided by Oded Magger.

4.5 Acknowledgments

The authors want to thank Jon Laurent and Kris McGary for some of the data used, and Li and Patra for making their code available. This work was supported by grants from the U.S. Army Research (58343- MA) to EMM and ISD, from the NSF, NIH, Welch (F1515) and Packard Foundations to EMM, and from DOD Army (W911NF-10-1-0529), NSF (CCF-0916309) and the Moncrief Grand Challenge

[2]from `http://cbio.ensmp.fr/~jvert/svn/prodige/html/prodige-0.3.tar.gz`
[3]from `http://marcottelab.org/index.php/Catapult/`
[4]from `http://www3.ntu.edu.sg/home/aspatra/research/Yongjin_BI2010.zip`
[5]from `http://cbio.ensmp.fr/~jvert/svn/prodige/html/prodige-0.3.tar.gz`

Table 4.3: **Different species used for inferring gene-phenotype associations in the proposed methods Katz and** CATAPULT**,** and sizes of the gene-phenotype networks for the species, restricted to orthologs of human genes. The total number of human genes with any kind of phenotype annotation is 12331.

Index	Species	# Phenotypes	# Associations
1	Human (*Hs*)	3,209	3,954
2	Plant (*At*)	1,137	12,010
3	Worm (*Ce*)	744	30,519
4	Fly (*Dm*)	2,503	68,525
5	Zebrafish (*Dr*)	1,143	4,500
6	*E.coli* (*Ec*)	324	72,846
7	Chicken (*Gg*)	1,188	22,150
8	Mouse (*Mm*)	4,662	75,199
9	Yeast (*Sc*)	1,243	73,284

Table 4.4: **Benchmark Drug data sets used for evaluation,** obtained from [197].

Index	Type	# Drugs	# Associations
1	Enzymes	445	2,926
2	Ion Channels	210	1,476
3	GPCRs	223	635
4	Nuclear Receptors	54	90

Award to ISD.

4.6 Supplementary Material

Relationship between Katz on the heterogenous network and RWRH

Restricting P to human phenotypes, i.e. letting $P = P_{Hs}$, and weighing P and $P^\top$ by λ in the heterogeneous network C where $0 < \lambda < 1$ is the jumping probability, in Equation (4.3), we get the heterogeneous network construction used in [173]. The random walk with restarts method of [173] when extended to our heterogenous network turns out to be *equivalent* to Katz measure provided the

Table 4.5: **Weights learned for different features by CATAPULT using the biased SVM with bagging procedure,** using the HumanNet gene network. Two important observations are: (1) Features corresponding to higher path lengths receive relatively much smaller weights. (2) Features corresponding to different species receive different weights, in particular, features derived from mouse phenotypes get the highest weights, which makes sense given the relative evolutionary proximity between humans and mice.

Type	Feature	Learnt weights	Feature	Learnt weights
Human	$P_{Hs}P_{Hs}^T P_{Hs}$	31.04	$P_{Hs}P_{Hs}^T G^2 P_{Hs}$	8.97
	$P_{Hs}P_{Hs}^T GP_{Hs}$	2.60	$GP_{Hs}P_{Hs}^T GP_{Hs}$	5.98
	$GP_{Hs}P_{Hs}^T P_{Hs}$	1.00	$G^2 P_{Hs}P_{Hs}^T P_{Hs}$	3.31
Plant	$P_{At}P_{At}^T P_{Hs}$	8.09	$P_{At}P_{At}^T G^2 P_{Hs}$	0.74
	$P_{At}P_{At}^T GP_{Hs}$	1.06	$GP_{At}P_{At}^T GP_{Hs}$	2.18
	$GP_{At}P_{At}^T P_{Hs}$	1.20	$G^2 P_{At}P_{At}^T P_{Hs}$	0.65
Worm	$P_{Ce}P_{Ce}^T P_{Hs}$	5.75	$P_{Ce}P_{Ce}^T G^2 P_{Hs}$	0.33
	$P_{Ce}P_{Ce}^T GP_{Hs}$	0.69	$GP_{Ce}P_{Ce}^T GP_{Hs}$	0.55
	$GP_{Ce}P_{Ce}^T P_{Hs}$	0.62	$G^2 P_{Ce}P_{Ce}^T P_{Hs}$	0.29
Fly	$P_{Dm}P_{Dm}^T P_{Hs}$	4.58	$P_{Dm}P_{Dm}^T G^2 P_{Hs}$	0.90
	$P_{Dm}P_{Dm}^T GP_{Hs}$	0.93	$GP_{Dm}P_{Dm}^T GP_{Hs}$	1.36
	$GP_{Dm}P_{Dm}^T P_{Hs}$	0.72	$G^2 P_{Dm}P_{Dm}^T P_{Hs}$	0.55
Zebrafish	$P_{Dr}P_{Dr}^T P_{Hs}$	8.28	$P_{Dr}P_{Dr}^T G^2 P_{Hs}$	1.16
	$P_{Dr}P_{Dr}^T GP_{Hs}$	0.77	$GP_{Dr}P_{Dr}^T GP_{Hs}$	2.68
	$GP_{Dr}P_{Dr}^T P_{Hs}$	0.52	$G^2 P_{Dr}P_{Dr}^T P_{Hs}$	0.69
E.coli	$P_{Ec}P_{Ec}^T P_{Hs}$	1.67	$P_{Ec}P_{Ec}^T G^2 P_{Hs}$	0.19
	$P_{Ec}P_{Ec}^T GP_{Hs}$	0.30	$GP_{Ec}P_{Ec}^T GP_{Hs}$	0.75
	$GP_{Ec}P_{Ec}^T P_{Hs}$	0.29	$G^2 P_{Ec}P_{Ec}^T P_{Hs}$	0.12
Chicken	$P_{Gg}P_{Gg}^T P_{Hs}$	3.76	$P_{Gg}P_{Gg}^T G^2 P_{Hs}$	0.32
	$P_{Gg}P_{Gg}^T GP_{Hs}$	0.30	$GP_{Gg}P_{Gg}^T GP_{Hs}$	1.35
	$GP_{Gg}P_{Gg}^T P_{Hs}$	0.23	$G^2 P_{Gg}P_{Gg}^T P_{Hs}$	1.82
Mouse	$P_{Mm}P_{Mm}^T P_{Hs}$	15.03	$P_{Mm}P_{Mm}^T G^2 P_{Hs}$	1.54
	$P_{Mm}P_{Mm}^T GP_{Hs}$	1.35	$GP_{Mm}P_{Mm}^T GP_{Hs}$	2.13
	$GP_{Mm}P_{Mm}^T P_{Hs}$	0.83	$G^2 P_{Mm}P_{Mm}^T P_{Hs}$	0.70
Yeast	$P_{Sc}P_{Sc}^T P_{Hs}$	5.55	$P_{Sc}P_{Sc}^T G^2 P_{Hs}$	0.30
	$P_{Sc}P_{Sc}^T GP_{Hs}$	0.61	$GP_{Sc}P_{Sc}^T GP_{Hs}$	0.59
	$GP_{Sc}P_{Sc}^T P_{Hs}$	0.56	$G^2 P_{Sc}P_{Sc}^T P_{Hs}$	0.25
Gene network	GP_{Hs}	1.23	$G^3 P_{Hs}$	0.57
	$G^2 P_{Hs}$	3.52	$G^4 P_{Hs}$	0.36
Phenotype network	$P_{Hs}Q_{Hs}$	39.63	$P_{Hs}P_{Hs}^T G^2 P_{Hs}Q_{Hs}$	4.28
	$P_{Hs}P_{Hs}^T P_{Hs}Q_{Hs}$	21.02	$GP_{Hs}P_{Hs}^T GP_{Hs}Q_{Hs}$	2.56
	$P_{Hs}P_{Hs}^T GP_{Hs}Q1$	1.70	$G^2 P_{Hs}P_{Hs}^T P_{Hs}Q_{Hs}$	1.43
	$GP_{Hs}P_{Hs}^T P_{Hs}Q_{Hs}$	0.64		

columns of the combined matrix C are normalized appropriately. The equivalence is shown below. Let C^N denote the *normalized* matrix, with the different blocks weighted as described above. Then,

$$C^N_{:,g} = \begin{bmatrix} \lambda \frac{G_{:,g}}{\|G_{:,g}\|_1} \\ (1-\lambda)\frac{P^\top_{g,:}}{\|P_{g,:}\|_1} \end{bmatrix}$$

where g refers to one of the first $|\mathcal{G}|$ columns of the matrix C, and

$$C^N_{:,p} = \begin{bmatrix} (1-\lambda)\frac{P^\top_{:,p}}{\|P_{:,p}\|_1} \\ \lambda \frac{Q^\top_{:,p}}{\|Q_{:,p}\|_1} \end{bmatrix}$$

where p refers to one of the remaining $|\mathcal{P}_{Hs}|$ columns of C, with the understanding that if a gene is not known to be associated to any phenotype (i.e. $\|P_{g,:}\| = 0$) then we will simply use $\lambda = 1$ for that gene. Case $\|Q_{:,p}\| = 0$ is handled similarly. Then we consider the evolution:

$$v_{T+1} = \beta C^N v_T + (1-\beta) C^N_{:,p}$$

where $C^N_{:,p}$ is simply a probability distribution with equal mass on all genes known to be associated with a phenotype p of interest, and mass on the diseases related to p . The genes are then ranked in the order of the mass that is assigned to them under the steady state distribution v_∞ of this evolution. The steady state vector v_∞ should satisfy

$$v_\infty = \beta C^N v_\infty + (1-\beta) C^N_{:,p}$$

which readily yields

$$v_\infty = (1-\beta)[I - \beta C^N]^{-1} C^N_{:,p} \, .$$

Thus the score matrix computed by RWRH can be written as[6],

$$\beta[I - \beta C^N]^{-1} C^N = \beta C^N + \beta^2 (C^N)^2 + \beta^3 (C^N)^3 + \dots .$$

which is *exactly* Katz but on the *normalized* matrix C^N instead of C itself.

Relationship between Katz on the heterogenous network and PRINCE

Examining the computation of Katz on heterogeneous network closely yields an interesting connection to PRINCE. As $k \to \infty$ in Equation (4.4) and for appropriate choice of β, let

$$S^{katz}(C) = (I - \beta C)^{-1} = \begin{bmatrix} S_{GG} & S_{GP} \\ (S_{GP})^\top & S_{PP} \end{bmatrix} ,$$

where it can be shown that

$$S_{GP} = S^{katz}(G) P \left[I - (Q + P^\top S^{katz}(G) P) \right]^{-1} . \tag{4.12}$$

Note how the Katz similarity matrix $S^{katz}(G) = (I - \beta G)^{-1}$ for the (weighted) gene-gene network $\lambda_G G$ itself appears in the expression above. The expression above takes into account all kinds of paths in the combined network that start in gene nodes and end up in human phenotype nodes. The corresponding score matrix computed by PRINCE[172] method can be generalized as

$$S_{GP}^{PRINCE} = S^{Katz}(G) P Q$$

Note that it is a form a generalization, as PRINCE smoothes a given phenotype using its most similar neighbor, whereas the term PQ in Equation (4.6) combines

[6]Multiplying either sides of the equation by constant factor $\beta/(1-\beta)$ does not affect the ranking of candidates.

all the neighbors linearly. Also note that the expression should strictly have P_{Hs} and Q_{Hs} instead of P and Q as [172] uses only human phenotypes data. However, writing this way enables comparison to the expression corresponding to Katz on heterogeneous network given in Equation (4.12). Clearly, Katz on heterogeneous network generalizes PRINCE method. In particular we observe that while PRINCE relies on the matrix Q to obtain "smoothed" phenotypes by sharing information across phenotypes, Katz on heterogeneous network uses a combination of Q and $P^\top S^{katz}(G)P$.

Chapter 5

Conclusions and musings about the future

The genomic era has completely revolutionized the study of the genetic basis for all kinds of human diseases and disorders. Cheap genotyping platforms, like SNP chips, have made it possible to find subtle genetic risk factors by genotyping huge populations and millions of commonly varying loci. At the other end of the spectrum, modern sequencing technology allows us to cheaply sequence candidate disease genes for very rare (or unique) mutations.

Yet, despite this deluge of data, the results of the genotyping efforts of the last decade have often met with disappointment. Genome-wide association studies have only found a small fraction of the genetic basis for common diseases, suggesting that much of the genetic risk lies in rarer mutations (the so-called "missing heritability"). Candidate gene sequencing studies often fail to find any interesting mutations, simply because the wrong candidate gene was sequenced.

In this thesis, I have used the information encoded by HumanNet to address both of these problems, to find genes weakly associated with diseases in GWAS in chapter 2 and to prioritize candidate genes to sequence for rare mutations in chapters 3 and 4. The "guilt-by-association" framework I have employed is not only a powerful tool for identifying new gene-disease associations, it also gives *context* to the predictions it generates; by studying the network neighbours of

the genes predicted and already associated with a disease, we get an idea of what pathways and biological processes are involved in the disease.

In the future, the distinction between the GWAS setting and the candidate gene sequencing setting will become weaker and weaker. With falling sequencing costs the need to use a SNP chip instead of fully sequencing the genomes of study participants will disappear. In the same way, the need to restrict sequencing to only a few candidate genes will disappear. It will therefore be necessary to let the methods for future GWAS analysis borrow from the tools we use to prioritize rare candidate genes, and to expand the techniques used to find rare mutations in Mendelian disorders so that they can be used to analyse genome-wide data. This work shows that network based GBA is a powerful tool at both ends of this spectrum, and will therefore be an important part in this future statistical framework.

Appendix A

Network-smoothed sparse regression for GWAS

A.1 Background

I have described the idea behind GWAS in the introduction. Briefly, at each SNP, a person can be either homozygous for the major allele, homozygous for the minor allele, or heterozygous. The idea underlying GWAS is that genetic loci that affect disease state will be distributed differently among the cases and among the controls, and by looking at the genes located close to these markers we can learn what causes the disease. In the traditional analysis of GWAS data, this difference has been measured by means of χ^2 tests or by Bayesian modeling, and a p-value or log-odds Bayes factor has been given for each SNP to indicate how likely it is to be involved in the disease. I will here employ a different approach, inspired by compressed sensing and ℓ_1 regularization. This approach was first described in [55]. I will give a brief overview of the approach, and then state the modifications I proposed to it, and hurdles I encountered. However, in the spring of 2012 I learned that Alexis Battle in Daphne Koller's group had already tried a *very* similar approach, and was very close to publishing, which made me abandon this line of inquiry.

If we let the number of subjects in the study be n, and the number of SNPs be N, we can encode this information as a big matrix A, where $\dim A = n \times N$, and

$a_{ij} \in \{-1, 0, +1\}$ depending on if subject i is homozygous (± 1) or heterozygous (0) at SNP j.

A.2 Linear regression

If we assume that the loci affect the phenotype in an additive, linear, manner we can model the phenotype value y_i of patient i as

$$y_i = \sum_j a_{ij} x_j + \mu + e_i,$$

or, in matrix form,

$$\mathbf{y} = A\mathbf{x} + \mu + \mathbf{e},$$

where x_j is effect size for locus j, μ is the population mean for the phenotype and $\mathbf{e}$ is non-genetic factors, which we will model as noise. If we assume that only a small number of loci (on the order of a few hundred) affect the disease, $\mathbf{x}$ is a sparse vector. We can now state this as a ℓ_1 minimization problem,

$$\min_{\mathbf{x},\mu} \quad \|\mathbf{x}\|_1 \tag{A.1}$$

$$\text{where} \quad \|A\mathbf{x} + \mu - \mathbf{y}\|_2 < \epsilon.$$

This is the lasso, and has been tried in the context of GWAS by Wu et al. in [55], with promising results.

A.3 Logistic regression

However, most GWAS don't deal with continuous phenotypes. Instead, they are of the case-control variety. For such settings, logistic regression might be

more appropriate. There we want to solve

$$\min_{\mathbf{x},\mu} \quad \sum_i \ln\left(1+\exp\left(y_i(A_{ij}x_j+\mu)\right)\right),$$

or, equivalently, maximize the likelihood of the data. We can give this a more compressed flavor by adding a regularizing term to it, which makes sense given our assumption that the data is sparse. With a regularizing term added, we get

$$\min_{\mathbf{x},\mu} \quad \sum_i \ln\left(1+\exp\left([\mathbf{y}^t(A\mathbf{x}+\mu)]_i\right)\right)+\lambda\|\mathbf{x}\|_1. \tag{A.2}$$

A.4 Network smoothing penalties

One way to improve the performance of these classifiers is to add more data of some form, which is the approach I (and independently Battle) settled on. One obvious candidate for a good data source would be a gene-gene interaction network, and then add a penalty when the total effect of all the SNPs associated with a gene differs to much from the total effect of SNPs associated with network neighbors of that SNP. This would give us an optimization problem of the form

$$\min_x \frac{1}{2}||Ax-y||_2^2+\lambda||x||_1+\frac{1}{2}x^T\Gamma x, \tag{A.3}$$

where x is the SNP effect size, y the measured phenotypic value, A the genotype matrix and Γ is a network laplacian derived from HumanNet, that we're smoothing over. If we expand the norm squared, we get

$$||Ax-y||_2^2+\beta x^T\Gamma x = x^TA^TAx-2yAx+y^2+x^T\Gamma x.$$

A.5 Approach

However, this drives the x-es to have the same sign if they are similar. In my application that doesn't make any sense, since the sign is an artifact of the encoding of the matrix A. I would therefore like to solve something like

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \lambda||x||_1 + \frac{1}{2}\sum_{ij}(|x_i| - |x_j|)^2 N_{ij}. \tag{A.4}$$

Clearly, this is not convex . However, I can find a convex relaxation. Consider the following rewrite of Eq. A.3:

$$\min_{x_+,x_-} \frac{1}{2}||A(x_+ - x_-) - y||_2^2 + \lambda||x_+ - x_-||_1 + \frac{1}{2}(x_+ - x_-)^T\Gamma(x_+ - x_-),$$

where x_+ and $x_- \geq 0$. The similarly rewritten version of Eq. A.4 is

$$\min_{x_+,x_-} \frac{1}{2}||A(x_+ - x_-) - y||_2^2 + \lambda||x_+ + x_-||_1 + \frac{1}{2}(x_+ + x_-)^T\Gamma(x_+ + x_-),$$

where

$$\begin{array}{llll} (i) & x_+ \geq 0, & (ii) & x_{+i} > 0 \Rightarrow x_{-i} = 0, \\ (iii) & x_- \geq 0, & (iv) & x_{-i} > 0 \Rightarrow x_{+i} = 0. \end{array}$$

If we relax the constraints by simply dropping conditions (ii) and (iv), we get a convex problem. Using the substitution $z = x_+ + x_-$ and $w = x_+ - x_-$, this can be written as

$$\min_{z,w} \frac{1}{2}||Aw - y||_2^2 + \lambda||z||_1 + \frac{1}{2}z^T\Gamma z,$$

where $|w_i| \leq z_i$, which is clearly convex.

A.6 Problem

The network smoothing seems like it should enforce the kind of structure I want. However, since it smooths over all values, nodes that are highly connected

will be pulled down by the zeros they're surrounded by. I'm not sure how to deal with that. One could normalize the links for all nodes. This breaks symmetry of the network, but maybe that's not a big deal. Another way could be to choose the penalties such that all nodes are a priori equally likely to be pulled down. That is, if all nodes are zero, the slope to increase a certain node is independent of the node. Since the slope at a node i is

$$\sum_j A_{ij}(A_{jk}w_k - y_j) + \lambda_i \text{sign}(z_i) + \Gamma_{ij}z_j$$

we could choose λ_i to be $\Lambda - \Gamma_{ii}$ or something like it.

While I was dealing with this problem, I saw a talk by Alexis Battle, and realized she already had a working network-smoothed GWAS regression model, and had tried it on the same data set I was planning to use.

Bibliography

[1] Lander, E. S., Linton, L. M., Birren, B. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

[2] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291**, 1304–51 (2001).

[3] The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–320 (2005).

[4] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78 (2007).

[5] Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics* **42**, 30–35 (2010).

[6] Buyse, I. M., Fang, P., Hoon, K. T., Amir, R. E. *et al.* Diagnostic Testing for Rett Syndrome by DHPLC and Direct Sequencing Analysis of the MECP2 Gene: Identification of Several Novel Mutations and Polymorphisms. *The American Journal of Human Genetics* **67**, 1428–1436 (2000).

[7] Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).

[8] Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. A probabilistic functional network of yeast genes. *Science (New York, N.Y.)* **306**, 1555–8 (2004).

[9] McGary, K. L., Lee, I. & Marcotte, E. M. Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. *Genome biology* **8**, R258 (2007).

[10] Lee, I., Lehner, B., Crombie, C., Wong, W. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nature genetics* **40**, 181–8 (2008).

[11] McGary, K. L., Park, T. J., Woods, J. O., Cha, H. J. *et al.* Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 6544–9 (2010).

[12] Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids. *Nature* **171**, 96–98 (1953).

[13] Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

[14] Levy, S., Sutton, G., Ng, P. C., Feuk, L. *et al.* The diploid genome sequence of an individual human. *PLoS biology* **5**, e254 (2007).

[15] Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature reviews Genetics* **6**, 95–108 (2005).

[16] Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *TRENDS in Genetics* **17**, 502–510 (2001).

[17] Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics* **69**, 124–37 (2001).

[18] Corvin, A., Craddock, N. & Sullivan, P. F. Genome-wide association studies: a primer. *Psychological medicine* **40**, 1063–77 (2010).

[19] Zhong, H., Yang, X., Kaplan, L. M., Molony, C. & Schadt, E. E. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *American journal of human genetics* **86**, 581–91 (2010).

[20] Manolio, T. a., Collins, F. S., Cox, N. J., Goldstein, D. B. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–53 (2009).

[21] Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research* **21**, 1109–21 (2011).

[22] Bonetta, L. Getting up close and personal with your genome. *Cell* **133**, 753–756 (2008).

[23] Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science (New York, N.Y.)* **322**, 881–8 (2008).

[24] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* **9**, 356–69 (2008).

[25] Liew, C.-C. & Dzau, V. J. Molecular genetics and genomics of heart failure. *Nature reviews. Genetics* **5**, 811–25 (2004).

[26] Pomp, D., Allan, M. F. & Wesolowski, S. R. Quantitative genomics: exploring the genetic architecture of complex trait predisposition. *Journal of animal science* **82 E-Suppl**, E300–E312 (2004).

[27] Visscher, P. M. Sizing up human height variation. (2008).

[28] Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H. *et al.* Many sequence variants affecting diversity of adult human height. *Nature genetics* **40**, 609–615 (2008).

[29] Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R. *et al.* Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature genetics* **40**, 584–591 (2008).

[30] Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics* **40**, 575–583 (2008).

[31] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42**, 565–9 (2010).

[32] Ideker, T. & Sharan, R. Protein networks in disease. *Genome research* **18**, 644–652 (2008).

[33] Christensen, C., Thakar, J. & Albert, R. Systems-level insights into cellular regulation: inferring, analysing, and modelling intracellular networks. *IET systems biology* **1**, 61–77 (2007).

[34] Lee, I., Narayanaswamy, R. & Marcotte, E. M. Bioinformatic prediction of yeast gene function. *Methods* **36**, 597–628 (2007).

[35] Bonneau, R. Learning biological networks: from modules to dynamics. *Nature chemical biology* **4**, 658–664 (2008).

[36] Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. O. Reconstruction of biochemical networks in microorganisms. *Nature reviews Microbiology* **7**, 129–143 (2009).

[37] Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8348–8353 (2003).

[38] Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V. *et al.* Probabilistic model of the human protein-protein interaction network. *Nature biotechnology* **23**, 951–959 (2005).

[39] Alexeyenko, A. & Sonnhammer, E. L. L. Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome research* **19**, 1107–1116 (2009).

[40] Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S. *et al.* STRING 8a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research* **37**, D412–D416 (2009).

[41] Fraser, H. B. & Plotkin, J. B. Using protein complexes to predict phenotypic effects of gene mutation. *Genome biology* **8**, R252 (2007).

[42] Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology* **25**, 309–16 (2007).

[43] Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* **37**, W305–W311 (2009).

[44] Huttenhower, C., Haley, E. M., Hibbs, M. A., Dumeaux, V. *et al.* Exploring the human genome with functional maps. *Genome Research* **19**, 1093–1106 (2009).

[45] Linghu, B., Snitkin, E. S., Hu, Z., Xia, Y. & DeLisi, C. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biology* **10**, R91 (2009).

[46] Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M. & Rhee, S. Y. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nature biotechnology* **28**, 149–156 (2010).

[47] Hart, G. T., Lee, I. & Marcotte, E. R. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC bioinformatics* **8**, 236 (2007).

[48] Emily, M., Mailund, T., Hein, J., Schauser, L. & Schierup, M. H. Using biological networks to search for interacting loci in genome-wide association studies. *European journal of human genetics EJHG* **17**, 1231–1240 (2009).

[49] Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J. *et al.* Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS genetics* **7**, 13 (2011).

[50] Hannum, G., Srivas, R., Guénolé, A., Van Attikum, H. *et al.* Genome-Wide Association Data Reveal a Global Map of Genetic Interactions among Protein Complexes. *PLoS genetics* **5**, 11 (2009).

[51] Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P. *et al.* Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human molecular genetics* **18**, 2078–2090 (2009).

[52] Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome biology* **11**, R53 (2010).

[53] Franke, L., Van Bakel, H., Fokkens, L., De Jong, E. D. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American journal of human genetics* **78**, 1011–1025 (2006).

[54] Pico, A. R., Smirnov, I. V., Chang, J. S. *et al.* SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system. *Nucleic acids research* **37**, D803–D809 (2009).

[55] Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. & Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics (Oxford, England)* **25**, 714–21 (2009).

[56] Chang, J. S., Yeh, R.-F., Wiencke, J. K., Wiemels, J. L. *et al.* Pathway analysis of single-nucleotide polymorphisms potentially associated with glioblastoma multiforme susceptibility using random forests. *Cancer epidemiology biomarkers prevention a publication of the American Association for Cancer Research cosponsored by the American Society of Preventive Oncology* **17**, 1368–1373 (2008).

[57] Saccone, S. F., Saccone, N. L., Swan, G. E., Madden, P. A. F. *et al.* Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics Oxford England* **24**, 1805–1811 (2008).

[58] Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics* **11**, 843–854 (2010).

[59] Lee, I., Li, Z. & Marcotte, E. M. An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. *PloS one* **2**, e988 (2007).

[60] Lee, I., Lehner, B., Vavouri, T., Shin, J. *et al.* Predicting genetic modifier loci using functional gene networks. *Genome research* **20**, 1143–53 (2010).

[61] Fishel, R., Lescoe, M. K., Rao, M. R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. (1994).

[62] Miyaki, M., Konishi, M., Tanaka, K., Kikuchi-Yanoshita, R. *et al.* Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nature genetics* **17**, 271–272 (1997).

[63] Moser, A. B., Rasmussen, M., Naidu, S., Watkins, P. A. *et al.* Phenotype of patients with peroxisomal disorders subdivided into sixteen complementation groups. (1995).

[64] Leegwater, P. a., Vermeulen, G., Könst, a. a., Naidu, S. *et al.* Subunits of the translation initiation factor eIF2B are mutant in leukoencephalopathy with vanishing white matter. *Nature genetics* **29**, 383–8 (2001).

[65] van der Knaap, M. S., Leegwater, P. A., Könst, A. A., Visser, A. *et al.* Mutations in each of the five subunits of translation initiation factor eIF2B can cause leukoencephalopathy with vanishing white matter. *Annals of neurology* **51**, 383–388 (2002).

[66] Schlabach, M. R., Luo, J., Solimini, N. L., Hu, G. *et al.* Cancer proliferation gene discovery through functional genomics. *Science New York NY* **319**, 620–4 (2008).

[67] Chang, K., Elledge, S. J. & Hannon, G. J. Lessons from Nature: microRNA-based shRNA libraries. *Nature methods* **3**, 707–714 (2006).

[68] Brass, A. L., Dykxhoorn, D. M., Benita, Y., Yan, N. *et al.* Identification of host proteins required for HIV infection through a functional genomic screen. *Science New York NY* **319**, 921–926 (2008).

[69] Ding, L., Paszkowski-Rogacz, M., Nitzsche, A., Slabicki, M. M. *et al.* A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. *Cell stem cell* **4**, 403–415 (2009).

[70] Luo, J., Emanuele, M. J., Li, D., Creighton, C. J. *et al.* A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* **137**, 835–848 (2009).

[71] Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Molecular systems biology* **3**, 88 (2007).

[72] Stolovitzky, G., Monroe, D. & Califano, A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Annals of the New York Academy of Sciences* **1115**, 1–22 (2007).

[73] Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology* **9 Suppl 1**, S4 (2008).

[74] Wang, P. I. & Marcotte, E. M. It's the machine that matters: Predicting gene function and phenotype from protein networks. *Journal of proteomics* **73**, 2277–2289 (2010).

[75] Liu, X., Yu, X., Zack, D. J., Zhu, H. & Qian, J. TiGER: A database for tissue-specific gene expression and regulation. *BMC bioinformatics* **9**, 271 (2008).

[76] Schwikowski, B., Uetz, P. & Fields, S. A network of protein-protein interactions in yeast. *Nature biotechnology* **18**, 1257–1261 (2000).

[77] Chen, J., Aronow, B. J. & Jegga, A. G. Disease candidate gene identification and prioritization using protein interaction networks. *BMC bioinformatics* **10**, 73 (2009).

[78] Ramakrishnan, S. R., Vogel, C., Kwon, T., Penalva, L. O. *et al.* Mining gene functional networks to improve mass-spectrometry-based protein identification. *Bioinformatics Oxford England* **25**, 2955–2961 (2009).

[79] Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).

[80] Suthram, S., Beyer, A., Karp, R. M., Eldar, Y. & Ideker, T. eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular systems biology* **4**, 162 (2008).

[81] Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E. & Blake, J. A. The Mouse Genome Database (MGD): from genes to micea community resource for mouse biology. *Nucleic acids research* **33**, D471–D475 (2005).

[82] Park, J.-H., Wacholder, S., Gail, M. H., Peters, U. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics* **42**, 570–5 (2010).

[83] Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS biology* **8**, e1000294 (2010).

[84] Wang, K., Dickson, S. P., Stolle, C. a., Krantz, I. D. *et al.* Interpretation of association signals and identification of causal variants from genome-wide association studies. *American journal of human genetics* **86**, 730–42 (2010).

[85] Jin, T. & Liu, L. The Wnt signaling pathway effector TCF7L2 and type 2 diabetes mellitus. *Molecular endocrinology (Baltimore, Md.)* **22**, 2383–92 (2008).

[86] Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics* **40**, 955–62 (2008).

[87] Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics* **40**, 638–45 (2008).

[88] Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nature protocols* **2**, 2366–2382 (2007).

[89] Van Limbergen, J., Wilson, D. C. & Satsangi, J. The genetics of Crohn's disease. *Annual review of genomics and human genetics* **10**, 89–116 (2009).

[90] Pai, R., Jones, M. K., Tomikawa, M. & Tarnawski, a. S. Activation of Raf-1 during experimental gastric ulcer healing is Ras-mediated and protein kinase C-independent. *The American journal of pathology* **155**, 1759–66 (1999).

[91] Hildt, E. & Oess, S. Identification of Grb2 as a novel binding partner of tumor necrosis factor (TNF) receptor I. *The Journal of experimental medicine* **189**, 1707–14 (1999).

[92] Reiley, W. W., Jin, W., Lee, A. J., Wright, A. *et al.* Deubiquitinating enzyme CYLD negatively regulates the ubiquitin-dependent kinase Tak1 and prevents abnormal T cell responses. *The Journal of experimental medicine* **204**, 1475–85 (2007).

[93] Regamey, A., Hohl, D., Liu, J. W. *et al.* The tumor suppressor CYLD interacts with TRIP and regulates negatively nuclear factor kappaB activation by tumor necrosis factor. *The Journal of experimental medicine* **198**, 1959–64 (2003).

[94] Beckly, J. B., Hancock, L., Geremia, A. *et al.* Two-stage candidate gene study of chromosome 3p demonstrates an association between nonsynonymous variants in the MST1R gene and Crohn's disease. *Inflammatory bowel diseases* **14**, 500–7 (2008).

[95] Hata, S., Abe, M., Suzuki, H., Kitamura, F. *et al.* Calpain 8/nCL-2 and calpain 9/nCL-4 constitute an active protease complex, G-calpain, involved in gastric mucosal defense. *PLoS genetics* **6**, e1001040 (2010).

[96] Cooper, J. D., Smyth, D. J., Smiles, A. M., Plagnol, V. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature genetics* **40**, 1399–401 (2008).

[97] Madu, I. G., Roth, S. L., Belouzard, S. & Whittaker, G. R. Characterization of a Highly Conserved Domain within the Severe Acute Respiratory Syndrome Coronavirus Spike Protein S2 Domain with Characteristics of a Viral Fusion Peptide. *Journal of virology* **83**, 7411–7421 (2009).

[98] Bonnefond, A., Froguel, P. & Vaxillaire, M. The emerging genetics of type 2 diabetes. *Trends in molecular medicine* **16**, 407–16 (2010).

[99] Koh, W., Mahan, R. D. & Davis, G. E. Cdc42- and Rac1-mediated endothelial lumen formation requires Pak2, Pak4 and Par3, and PKC-dependent signaling. *Journal of cell science* **121**, 989–1001 (2008).

[100] Li, Y., Sun, H., Wu, G., Du, W. *et al.* Protein kinase C/zeta (PRKCZ) gene is associated with type 2 diabetes in Han population of North China and analysis of its haplotypes. *World J Gastroenterol* **9(9)**, 2078–82 (2003).

[101] Akerblad, P., Månsson, R., Lagergren, A., Westerlund, S. *et al.* Gene expression analysis suggests that EBF-1 and PPARgamma2 induce adipogenesis of NIH-3T3 cells with similar efficiency and kinetics. *Physiological genomics* **23**, 206–16 (2005).

[102] Flint, J. & Mackay, T. F. C. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome research* **19**, 723–733 (2009).

[103] Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–8 (2001).

[104] Tong, A. H. Y., Lesage, G., Bader, G. D., Ding, H. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303**, 808–13 (2004).

[105] Davierwala, A. P., Haynes, J., Li, Z., Brost, R. L. *et al.* The synthetic genetic interaction spectrum of essential genes. *Nature genetics* **37**, 1147–1152 (2005).

[106] Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y. *et al.* The genetic landscape of a cell. *Science New York NY* **327**, 425–31 (2010).

[107] Lehner, B., Crombie, C., Tischler, J., Fortunato, A. & Fraser, A. G. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nature genetics* **38**, 896–903 (2006).

[108] Byrne, A. B., Weirauch, M. T., Wong, V. *et al.* A global analysis of genetic interactions in Caenorhabditis elegans. *Journal of biology* **6**, 8 (2007).

[109] Van Dongen, S. A cluster algorithm for graphs. Tech. Rep. 10, National Research Institute for Mathematics and Computer Science in the Netherlands (2000).

[110] Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* **30**, 1575–1584 (2002).

[111] Shubin, N., Tabin, C. & Carroll, S. Fossils, genes and the evolution of animal limbs. *Nature* **388**, 639–48 (1997).

[112] Ostlund, G., Schmitt, T., Forslund, K., Köstler, T. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic acids research* **38**, D196–203 (2010).

[113] Wang, X., Sun, W., Zhu, X., Li, L. *et al.* Association between the gamma-aminobutyric acid type B receptor 1 and 2 gene polymorphisms and mesial temporal lobe epilepsy in a Han Chinese population. *Epilepsy research* **81**, 198–203 (2008).

[114] Glasscock, E., Yoo, J. W., Chen, T. T., Klassen, T. L. & Noebels, J. L. Kv1.1 potassium channel deficiency reveals brain-driven cardiac dysfunction as a candidate mechanism for sudden unexplained death in epilepsy. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **30**, 5167–75 (2010).

[115] Butt, S. J. B., Sousa, V. H., Fuccillo, M. V., Hjerling-Leffler, J. *et al.* The requirement of Nkx2-1 in the temporal specification of cortical interneuron subtypes. *Neuron* **59**, 722–32 (2008).

[116] Ratnapriya, R., Vijai, J., Kadandale, J. S. *et al.* A locus for juvenile myoclonic epilepsy maps to 2q33-q36. *Human genetics* **128**, 123–30 (2010).

[117] Wyneken, U., Smalla, K. H., Marengo, J. J., Soto, D. *et al.* Kainate-induced seizures alter protein composition and N-methyl-D-aspartate receptor function of rat forebrain postsynaptic densities. *Neuroscience* **102**, 65–74 (2001).

[118] Douaud, M., Feve, K., Pituello, F., Gourichon, D. *et al.* Epilepsy Caused by an Abnormal Alternative Splicing with Dosage Effect of the SV2A Gene in a Chicken Model. *PloS one* **6**, e26932 (2011).

[119] Kaminski, R. M., Gillard, M., Leclercq, K., Hanon, E. *et al.* Proepileptic phenotype of SV2A-deficient mice is associated with reduced anticonvulsant efficacy of levetiracetam. *Epilepsia* **50**, 1729–40 (2009).

[120] Janz, R., Goda, Y., Geppert, M., Missler, M. & Südhof, T. C. SV2A and SV2B function as redundant Ca2+ regulators in neurotransmitter release. *Neuron* **24**, 1003–16 (1999).

[121] Crowder, K. M., Gunther, J. M., Jones, T. a., Hale, B. D. *et al.* Abnormal neurotransmission in mice lacking synaptic vesicle protein 2A (SV2A). *Proceedings of the National Academy of Sciences of the United States of America* **96**, 15268–73 (1999).

[122] Bagetta, G., Corasaniti, M. T., Iannone, M., Nisticò, G. & Stephenson, J. D. Production of limbic motor seizures and brain damage by systemic and intracerebral injections of paraquat in rats. *Pharmacology & toxicology* **71**, 443–8 (1992).

[123] De Sarro, A., Ammendola, D., Zappala, M., Grasso, S. & De Sarro, G. B. Relationship between structure and convulsant properties of some beta-lactam antibiotics following intracerebroventricular microinjection in rats. *Antimicrobial agents and chemotherapy* **39**, 232–7 (1995).

[124] Barger, G. & Dale, H. H. Chemical structure and sympathomimetic action of amines. *The Journal of physiology* **41**, 19–59 (1910).

[125] Skovgaard, N., Møller, K., Gesser, H. & Wang, T. Histamine induces postprandial tachycardia through a direct effect on cardiac H2-receptors in pythons. *American journal of physiology. Regulatory, integrative and comparative physiology* **296**, R774–85 (2009).

[126] Shigenobu, K., Tatsuno, H., Matsuki, N., Oshima, T. & Kasuya, Y. Electrophysiological and mechanical studies on the cardiac effects of a histamine H2 receptor antagonist, cimetidine, in the isolated guinea pig myocardium. *J Pharm Dyn* **2**, 141–150 (1979).

[127] Levi, R., Malm, J. R., Bowman, F. O. & Rosen, M. R. The arrhythmogenic actions of histamine on human atrial fibers. *Circulation research* **49**, 545–50 (1981).

[128] He, G., Hu, J., Li, T., Ma, X. *et al.* The arrhythmogenic effect of sympathetic histamine in mouse hearts subjected to acute ischemia. *Molecular medicine (Cambridge, Mass.)* (2011).

[129] Hammadi, M., Adi, M., John, R., Khoder, G. A. K. & Karam, S. M. Dysregulation of gastric H,K-ATPase by cigarette smoke extract. *World journal of gastroenterology : WJG* **15**, 4016–22 (2009).

[130] Trivedi, C. M., Cappola, T. P., Margulies, K. B. & Epstein, J. a. Homeodomain only protein x is down-regulated in human heart failure. *Journal of molecular and cellular cardiology* **50**, 1056–8 (2011).

[131] Ellinor, P. T., Petrov-Kondratov, V. I., Zakharova, E., Nam, E. G. & MacRae, C. a. Potassium channel gene mutations rarely cause atrial fibrillation. *BMC medical genetics* **7**, 70 (2006).

[132] Xu, L.-X., Yang, W.-Y., Zhang, H.-Q., Tao, Z.-H. & Duan, C.-C. [Study on the correlation between CETP TaqIB, KCNE1 S38G and eNOS T-786C gene polymorphisms for predisposition and non-valvular atrial fibrillation]. *Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi* **29**, 486–92 (2008).

[133] Yao, J., Ma, Y.-t., Xie, X., Liu, F. *et al.* [Association of rs1805127 polymorphism of KCNE1 gene with atrial fibrillation in Uigur population of Xinjiang]. *Zhonghua yi xue yi chuan xue za zhi = Zhonghua yixue yichuanxue zazhi = Chinese journal of medical genetics* **28**, 436–40 (2011).

[134] Beyer, E. C., Paul, D. L. & Goodenough, D. a. Connexin43: a protein from rat heart homologous to a gap junction protein from liver. *The Journal of cell biology* **105**, 2621–9 (1987).

[135] Delorme, B., Dahl, E., Jarry-Guichard, T., Marics, I. *et al.* Developmental regulation of connexin 40 gene expression in mouse heart correlates with the differentiation of the conduction system. *Developmental dynamics : an official publication of the American Association of Anatomists* **204**, 358–71 (1995).

[136] Lin, X., Gemel, J., Glass, A., Zemlin, C. W. *et al.* Connexin40 and connexin43 determine gating

properties of atrial gap junction channels. *Journal of molecular and cellular cardiology* **48**, 238–45 (2010).

[137] Cottrell, G. T. & Burt, J. M. Heterotypic gap junction channel formation between heteromeric and homomeric Cx40 and Cx43 connexons. *American journal of physiology. Cell physiology* **281**, C1559–67 (2001).

[138] Reaume, A. G., de Sousa, P. A., Kulkarni, S., Langille, B. L. *et al.* Cardiac malformation in neonatal mice lacking connexin43. *Science (New York, N.Y.)* **267**, 1831–4 (1995).

[139] Tuomi, J. M., Tyml, K. & Jones, D. L. Atrial tachycardia/fibrillation in the connexin 43 G60S mutant (Oculodentodigital dysplasia) mouse. *American journal of physiology. Heart and circulatory physiology* **300**, H1402–11 (2011).

[140] Thibodeau, I. L., Xu, J., Li, Q., Liu, G. *et al.* Paradigm of genetic mosaicism and lone atrial fibrillation: physiological characterization of a connexin 43-deletion mutant identified from atrial tissue. *Circulation* **122**, 236–44 (2010).

[141] Laitinen-Forsblom, P. J., Mäkynen, P., Mäkynen, H., Yli-Mäyry, S. *et al.* SCN5A mutation associated with cardiac conduction defect and atrial arrhythmias. *Journal of cardiovascular electrophysiology* **17**, 480–5 (2006).

[142] Darbar, D., Kannankeril, P. J., Donahue, B. S., Kucera, G. *et al.* Cardiac sodium channel (SCN5A) variants associated with atrial fibrillation. *Circulation* **117**, 1927–35 (2008).

[143] Chen, L., Zhang, W., Fang, C., Jiang, S. *et al.* Polymorphism H558R in the Human Cardiac Sodium Channel SCN5A Gene is Associated with Atrial Fibrillation. *The Journal of international medical research* **39**, 1908–16 (2011).

[144] Fraser, G. R., Froggatt, P. & James, T. N. Congenital deafness associated with electrocardiographic abnormalities, fainting attacks and sudden death. A recessive syndrome. *The Quarterly journal of medicine* **33**, 361–85 (1964).

[145] Fraser, G. R., Froggatt, P. & Murphy, T. Genetical aspects of the cardio-auditory syndrome of Jervell and Lange-Nielsen (Congenital deafness and electrocardiographic abnormalities). *Annals of human genetics* **28**, 133–57 (1964).

[146] Schwartz, P. J. The long QT syndrome. *Current problems in cardiology* **22**, 297–351 (1997).

[147] Schwartz, P. J., Spazzolini, C., Crotti, L., Bathen, J. *et al.* The Jervell and Lange-Nielsen syndrome: natural history, molecular basis, and clinical outcome. *Circulation* **113**, 783–90 (2006).

[148] Neyroud, N., Tesson, F., Denjoy, I., Leibovici, M. *et al.* A novel mutation in the potassium channel gene KVLQT1 causes the Jervell and Lange-Nielsen cardioauditory syndrome. *Nature genetics* **15**, 186–9 (1997).

[149] Schulze-Bahr, E., Wang, Q., Wedekind, H., Haverkamp, W. *et al.* KCNE1 mutations cause Jervell and Lange-Nielsen syndrome. *Nature genetics* **17**, 267–8 (1997).

[150] Gritli, S., Ben Salah, M., Shili, A., Robson, C. D. *et al.* Association of the long QT syndrome With goiter and deafness. *The American journal of cardiology* **105**, 681–6 (2010).

[151] Belmont, J. W., Craigen, W., Martinez, H. & Jefferies, J. L. Genetic disorders with both hearing loss and cardiovascular abnormalities. *Advances in oto-rhino-laryngology* **70**, 66–74 (2011).

[152] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).

[153] Chanvivattana, Y., Bishopp, A., Schubert, D., Stock, C. *et al.* Interaction of Polycomb-group proteins controlling flowering in Arabidopsis. *Development (Cambridge, England)* **131**, 5263–76 (2004).

[154] Koornneef, M., Hanhart, C. J. & van der Veen, J. H. A genetic and physiological analysis of late flowering mutants in Arabidopsis thaliana. *Molecular & general genetics : MGG* **229**, 57–66 (1991).

[155] Genger, R. K., Peacock, W. J., Dennis, E. S. & Finnegan, E. J. Opposing effects of reduced DNA methylation on flowering time in Arabidopsis thaliana. *Planta* **216**, 461–6 (2003).

[156] Weigel, D., Ahn, J. H., Blázquez, M. A., Borevitz, J. O. *et al.* Activation tagging in Arabidopsis. *Plant physiology* **122**, 1003–13 (2000).

[157] Abe, M., Katsumata, H., Komeda, Y. & Takahashi, T. Regulation of shoot epidermal cell differentiation by a pair of homeodomain proteins in Arabidopsis. *Development (Cambridge, England)* **130**, 635–43 (2003).

[158] Ikeda, Y., Kobayashi, Y., Yamaguchi, A., Abe, M. & Araki, T. Molecular basis of late-flowering phenotype caused by dominant epi-alleles of the FWA locus in Arabidopsis. *Plant & cell physiology* **48**, 205–20 (2007).

[159] Green, R. a., Kao, H.-L., Audhya, A., Arur, S. *et al.* A High-Resolution C.elegans Essential Gene Network Based on Phenotypic Profiling of a Complex Tissue. *Cell* **145**, 470–482 (2011).

[160] Klein, T. E., Chang, J. T., Cho, M. K., Easton, K. L. *et al.* Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *The pharmacogenomics journal* **1**, 167–70 (2001).

[161] Nichols, R. J., Sen, S., Choo, Y. J., Beltrao, P. *et al.* Phenotypic landscape of a bacterial cell. *Cell* **144**, 143–56 (2011).

[162] Tweedie, S., Ashburner, M., Falls, K. *et al.* FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic acids research* **37**, D555–9 (2009).

[163] Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T. *et al.* The Zebrafish Information Network: the zebrafish model organism database. *Nucleic acids research* **34**, D581–5 (2006).

[164] Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z. *et al.* The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic acids research* **36**, D1009–14 (2008).

[165] Bell, G. W., Yatskievych, T. a. & Antin, P. B. GEISHA, a whole-mount in situ hybridization gene expression screen in chicken embryos. *Developmental dynamics : an official publication of the American Association of Anatomists* **229**, 677–87 (2004).

[166] Goh, K., Cusick, M., Valle, D., Childs, B. *et al.* The human disease network. *Proceedings of the National Academy of Sciences* **104**, 8685 (2007).

[167] Tian, W., Zhang, L. V., Taan, M., Gibbons, F. D. *et al.* Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. *Genome Biology* **9 Suppl 1**, S7 (2008).

[168] Ulitsky, I. & Shamir, R. Identification of functional modules using network topology and high-throughput data. *BMC systems biology* **1**, 8 (2007).

[169] Human Protein Reaction Database, HPRD (2012). URL `http://www.hprd.org`.

[170] Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Mol Syst Biol* **4**, 189 (2008).

[171] Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics* 949–958 (2008).

[172] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology* **6**, e1000641 (2010).

[173] Li, Y. & Patra, J. C. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics/computer Applications in The Biosciences* **26**, 1219–1224 (2010).

[174] Mordelet, F. & Vert, J.-P. Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics* **12** (2011).

[175] Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).

[176] Gillis, J. & Pavlidis, P. The impact of multifunctional genes on "guilt by association" analysis. *PloS one* **6**, e17258 (2011).

[177] Cheng, F. *et al.* Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. *PLoS Comput Biol* **8**, e1002503+ (2012). URL `http://dx.doi.org/10.1371/journal.pcbi.1002503`.

[178] Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* **58**, 1019–1031 (2007).

[179] Estrada, E. & Higham, D. J. Network properties revealed through matrix functions. *SIAM Rev.* **52**, 696–714 (2010).

[180] Page, L., Brin, S., Motwani, R. & Winograd, T. The pagerank citation ranking: Bringing order to the web. (1999).

[181] Lu, Z., Savas, B., Tang, W. & Dhillon, I. Supervised link prediction using multiple sources. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, 923–928 (IEEE, 2010).

[182] Liu, B., Dai, Y., Li, X., Lee, W. S. & Yu, P. S. Building text classifiers using positive and unlabeled examples. In *In: Intl. Conf. on Data Mining*, 179–188 (2003).

[183] Mordelet, F. & Vert, J.-P. A bagging SVM to learn from positive and unlabeled examples. Tech. Rep. hal-00523336, version 1, HAL (2010).

[184] Lee, W. S. & Liu, B. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)* (2003).

[185] Online Mendelian Inheritance in Man, OMIM (2011). URL `http://omim.org/`.

[186] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research* **13**, 2363–71 (2003).

[187] Karni, S., Soreq, H. & Sharan, R. A network-based method for predicting disease-causing genes. *Journal of Computational Biology* **16**, 181–189 (2009).

[188] Chen, N., Harris, T. W., Antoshechkin, I. *et al.* WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic acids research* **33**, D383–9 (2005).

[189] Tweedie, S., Ashburner, M., Falls, K. *et al.* FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic acids research* **37**, D555–9 (2009).

[190] Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A. & Richardson, J. E. The mouse genome database (MGD): new features facilitating a model system. *Nucleic acids research* **35**, D630–7 (2007).

[191] Dwight, S. S., Harris, M. a., Dolinski, K. *et al.* Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic acids research* **30**, 69–72 (2002).

[192] Saito, T. L., Ohtani, M., Sawai, H., Sano, F. *et al.* SCMD: Saccharomyces cerevisiae Morphological Database. *Nucleic acids research* **32**, D319–22 (2004).

[193] Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E. *et al.* The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (New York, N.Y.)* **320**, 362–5 (2008).

[194] Sprague, J., Bayraktaroglu, L., Clements, D. *et al.* The Zebrafish Information Network: the zebrafish model organism database. *Nucleic acids research* **34**, D581–5 (2006).

[195] Bell, G. W., Yatskievych, T. a. & Antin, P. B. GEISHA, a whole-mount in situ hybridization gene expression screen in chicken embryos. *Developmental dynamics : an official publication of the American Association of Anatomists* **229**, 677–87 (2004).

[196] Van Driel, M., Bruggeman, J., Vriend, G., Brunner, H. & Leunissen, J. A text-mining analysis of the human phenome. *European journal of human genetics* **14**, 535–542 (2006).

[197] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. & Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. In *ISMB*, 232–240 (2008).

[198] Molecular Modeling and Design, LMMD (2012). URL `http://www.lmmd.org/database/dti`.